Nonparametric method for sparse conditional density estimation in moderately large dimensions

Minh-Lien Jeanne Nguyen

Nonparametric method for sparse conditional density estimation in moderately large dimensions

Minh-Lien Jeanne Nguyen

Laboratoire de Mathématiques d'Orsay
Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France

January 19, 2018

Abstract: In this paper, we consider the problem of estimating a conditional density in moderately large dimensions. Much more informative than regression functions, conditional densities are of main interest in recent methods, particularly in the Bayesian framework (studying the posterior distribution, finding its modes...). Considering a recently studied family of kernel estimators, we select a pointwise multivariate bandwidth by revisiting the greedy algorithm Rodeo (Regularisation Of Derivative Expectation Operator). The method addresses several issues: being greedy and computationally efficient by an iterative procedure, avoiding the curse of high dimensionality under some suitably defined sparsity conditions by early variable selection during the procedure, converging at a quasi-optimal minimax rate.

Keywords: *conditional density, high dimension, minimax rates, kernel density estimators, greedy algorithm, sparsity, nonparametric inference.*

1 Introduction

1.1 Motivations

In this paper, we consider the problem of the conditional density estimation. We observe a n-sample of a couple (X, Y), in which Y is the vector of interest while X gathers auxiliary variables. We denote d the joint dimension. In particular we are interested in the inference of the d-dimensional conditional density f of Y conditionally to X.

There is a growing demand for methods of conditional density estimation in a wide spectrum of applications such as Economy[Hall et al. 2004], Cosmology [Izbicki and Lee 2016], Medicine [Takeuchi et al. 2009], Actuaries [Efromovich 2010b], Meteorology [Jeon and Taylor 2012] among others. It can be explained by the double role of the conditional density estimation: deriving the underlying distribution of a dataset and determining the impact of the vector X of auxiliary variables on the vector of interest Y. In this aspect, the conditional density estimation is richer than both the unconditional density estimation and the regression problem. In particular, in the regression framework, only the conditional mean $\mathbb{E}[Y|X]$ are estimated instead of the full conditional density, which can be especially poorly informative in case of an asymmetric or multi-modal conditional density. Conversely, from the conditional density estimators, one can, *e.g.*, derive the conditional quantiles [Takeuchi et al. 2006] or give accurate predictive intervals [Fernández-Soto et al. 2002]. Furthermore, since the posterior distribution in the Bayesian framework is actually a conditional density, the present paper also offers an alternative method to the ABC methodology (for Approximate Bayesian Computation) [Beaumont et al. 2002 ; Marin et al. 2012 ; Biau et al. 2015] in the case of an intractable-yet-simulable model.

The challenging issue in conditional density estimation is to circumvent the "curse of dimensionality". The problem is twofold: theoretical and practical. In theory, it is stigmatized by the minimax approach, stating that in a d-dimensional space the best convergence rate for the pointwise risk over a p-regular class of functions is $\mathcal{O}(n^{-\frac{p}{2p+d}})$: in particular, the larger is d, the slower is the rate. In practice, the larger the dimension is, the larger the sample size is needed to control the estimation error. In order to maintain reasonable running times

in moderately large dimensions, methods have to be designed especially greedy.
Furthermore, one interesting question is how to retrieve the eventual *relevant* components in case of sparsity structure on the conditional density f. For example, if we have at disposal plenty of auxiliary variables without any indication on their dependency with our vector of interest Y, the ideal procedure will take in input the whole dataset and still achieve a running time and a minimax rate as fast as if only the relevant components were given and considered for the estimation. More precisely, two goals are simultaneously addressed : converging at rate $\mathcal{O}(n^{-\frac{2p}{2p+r}})$ with r the relevant dimension, *i.e.* the number of components that influence the conditional density f, and detect the irrelevant components at an early stage of the procedure in order to afterwards only work on the relevant data and thus speed up the running time.

1.2 Existing methodologies

Several nonparametric methods have been proposed to estimate conditional densities: kernel density estimators [Rosenblatt 1969 ; Hyndman et al. 1996 ; Bertin et al. 2016] and various methodologies for the selection of the associated bandwidth [Bashtannyk and Hyndman 2001 ; Fan and Yim 2004 ; Hall et al. 2004]; local polynomial estimators [Fan et al. 1996 ; Hyndman and Yao 2002]; projection series estimators [Efromovich 1999; 2007]; piecewise constant estimator [Györfi and Kohler 2007 ; Sart 2017]; copula [Faugeras 2009]. But while most of the aforementioned works are only defined for bivariate data or at least when either X or Y is univariate, they are also computationally intractable as soon as $d > 3$.
It is in particular the case for the kernel density methodologies (Hall, Racine ,Li 2004, Bertin et al. 2016): they achieve the optimal minimax rate, and even the detection of the relevant components, thanks to an adequate choice of the bandwidth (for the two aforementioned methods by cross validation and Goldenshluger-Lepski methodology), but the computational cost of these bandwidth selections is prohibitive even for moderate sizes of n and d. To the best of our knowledge, only two kernel density methods have been proposed to handle large datasets. [Holmes et al. 2010] propose a fast method of approximated cross-validation, based on a dual-tree speed-up, but they do not establish any rate of convergence and only show the consistency of their method. For scalar Y, [Fan et al. 2009] proposed to perform a prior step of dimension reduction on X to bypass the curse of dimensionality, then they estimate the bivariate approximated conditional density by kernel estimators. But the proved convergence rate $n^{-\frac{1}{3}}$ is not the optimal minimax rate $n^{-\frac{3}{8}}$ for the estimation of a bivariate function of assumed regularity 3. Moreover, the step of dimension reduction restricts the dependency of X to a linear combination of its components, which may induce a significant loss of information.
Projection series methods for scalar Y have also been proposed. [Efromovich 2010a] extends his previous work [Efromovich 2007] to a multivariate X. Theoretically the method achieves an oracle inequality, thus the optimal minimax rate. Moreover it performs an automatic dimension reduction on X when there exists a smaller intrinsic dimension. To our knowledge, it is the only method which addresses datasets of dimension larger than 3 with reasonable running times and does not pay its numerical performance with non optimal minimax rates. However the computation cost is prohibitive when both n and d are large. More recently, Izbicki and Lee have proposed two methodologies using orthogonal series estimators [Izbicki and Lee 2016; 2017]. The first method is particularly fast and can handle very large X (with more than 1000 covariates). Moreover the convergence rate adapts to an eventual smaller unknown intrinsic dimension of the support of the conditional density. The second method originally proposes to convert successful high dimensional regression methods into the conditional density estimation, interpreting the coefficients of the orthogonal series estimator as regression functions, which allows to adapt to all kind of figures (mixed data, smaller intrinsic dimension, relevant variables) in function of the regression method. However both methods converge slower than the optimal minimax rate. Moreover their optimal tunings depend in fact on the unknown intrinsic dimension.
For multivariate X and Y, [Otneim and Tjøstheim 2017] propose a new semiparametric method, called Locally Gaussian Density Estimator: they rewrite the conditional density as a product of a function depending on the marginal distribution functions (easily estimated since univariate, then plug-in), and a term which measures the dependency between the components, which is approximated by a centred Gaussian whose covariance is parametrically estimated. Numerically, the methodology seems robust to addition of covariates of X independent of Y, but it is not proved. Moreover they only establish the asymptotic normality of their method.

1.3 Our strategy and contributions

The challenge in this paper is to handle large datasets, thus we assume at our disposal a sample of large size n and of moderately large dimension. Then our work is motivated by the following three objectives:

(i) achieving the optimal minimax rate (up to a logarithm term);

(ii) being greedy, meaning that the procedure must have reasonable running times for large n and moderately large dimensions, in particular when $d > 3$;

(iii) adapting to a potential sparsity structure of f. More precisely, in the case where f locally depends only on a number r of its d components, r can be seen as the local *relevant* dimension. Then the desired convergence rate has to adapt to the unknown relevant dimension r: under this sparsity assumption, the benchmark for the estimation of a p-regular function is to achieve a convergence rate of the order $\mathcal{O}(n^{-\frac{2p}{2p+r}})$, which is the optimal minimax rate if the relevant components were given by an oracle.

Our strategy is based on kernel density estimators. The considered family has been recently introduced and studied in [Bertin et al. 2016]. This family is especially designed for conditional densities and is better adapted for the objective (iii) than the intensively studied estimator built as the ratio of a kernel estimator of the joint density over one of the marginal density of X. For example, a relevant component for the joint density and the marginal density of X may be irrelevant for the conditional density and it is the case if a component of X is independent of Y. Note though that many more cases of irrelevance exist since we define the relevance as a local property.
The main issue with kernel density estimators is the selection of the bandwidth $h \in \mathbb{R}_+^d$, and in our case, we also want to complete the objective (ii), since the pre-existing methodologies of bandwidth selection does not satisfy this restriction and thus cannot handle large datasets. In this paper, it is performed by an algorithm we call CDRodeo, which is derived from the algorithm Rodeo [Lafferty and Wasserman 2008 ; Liu et al. 2007], which has respectively been applied for the regression and the unconditional density estimation. The greediness of the algorithm allows us to address datasets of large sizes while keeping a reasonable running time (see Section 3.5 for further details). We give a simulated example with a sample of size $n = 10^5$ and of dimension $d = 5$ in Section 4. Moreover, Rodeo-type algorithms ensure an early detection of irrelevant component, and thus achieve the objective (iii) while improving the objective (ii).
From the theoretical point of view, if the regularity of f is known, our method achieves an optimal minimax rate (up to a logarithmic factor), which is adaptive to the unknown sparsity of f. The last property is mostly due to the Rodeo-type procedures. The improvement of our method in comparison to the paper [Liu et al. 2007] which estimates the *unconditional* density with Rodeo is twofold. First, our result is extended to any regularity $p \in \mathbb{N}_{>0}$, whereas [Liu et al. 2007] fixed $p = 2$. Secondly, our notion of relevance is both less restrictive and more natural. In [Liu et al. 2007], they studied the L_2-risk of their estimator, therefore they have to consider a notion of global relevance, whereas we consider a pointwise approach, which allows us to define a local property of relevance, which can be applied to a broader class of functions. Moreover, their notion of relevance is not intrinsic to the unknown density, but in fact depends on a tuning of the method, a prior chosen *baseline density* which has no connexion with the density, which limits the interpretation of the *relevance*.

1.4 Overview

Our paper is organized as follows. We introduce the CDRodeo method in Section 2. The theoretical results are in Section 3, in which we specify the assumptions and the tunings of the procedure from which are derived the convergence rate and the complexity cost of the method. A numerical example is presented in Section 4. The proofs are in the last section.

2 CDRodeo method

Let $W_1, \ldots, W_n$ be a sample of a couple (X, Y) of multivariate random vectors: for $i = 1, \ldots, n$,

$$W_i = (X_i, Y_i),$$

with X_i valued in $\mathbb{R}^{d_1}$ and Y_i in $\mathbb{R}^{d_2}$. We denote $d := d_1 + d_2$ the joint dimension.
We assume that the marginal distribution of X and the conditional distribution of Y given X are absolutely continuous with respect to the Lebesgue measure, and we define $f : \mathbb{R}^d \to \mathbb{R}$ such as for any $x \in \mathbb{R}^{d_1}$, $f(x, \cdot)$ is the conditional density of Y conditionally to $X = x$. We denote f_X the marginal density of X.
Our method estimates f pointwisely : let us fix $w = (x, y) \in \mathbb{R}^d$ the point of interest.

Kernel estimators. Our method is based on kernel density estimators. More specifically, we consider the family proposed in [Bertin et al. 2016], which is especially designed for the conditional density estimation. Let $K : \mathbb{R} \to \mathbb{R}$ be a kernel function, *ie:* $\int_{\mathbb{R}} K(t)dt = 1$, then for any bandwidth $h \in (\mathbb{R}_+^*)^d$, the estimator of $f(w)$ is defined by:

$$\hat{f}_h(w) := \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\tilde{f}_X(X_i)} \prod_{j=1}^{d} h_j^{-1} K\left(\frac{w_j - W_{ij}}{h_j}\right), \tag{1}$$

where $\tilde{f}_X$ is an estimator of f_X, built from another sample $\widetilde{X}$ of X. We denote by n_X the sample size of $\widetilde{X}$. The choices of K and $\tilde{f}_X$ are specified later (see section 3.2).

Bandwidth selection. In kernel density estimation, selecting the bandwidth is a critical choice which can be viewed as a bias-variance trade-off. In [Bertin et al. 2016], it is performed by the Goldenshluger-Lepski methodology (see [Goldenshluger and Lepski 2011]) and requires an optimization over an exhaustive grid of couples (h, h') of bandwidths, which leads to intractable running time when the dimension exceeds 3 (and large dataset).
That is why we focus in a method which excludes optimization over an exhaustive grid of bandwidths to rather propose a greedy algorithm derived from the algorithm Rodeo. First introduced in the regression framework [Wasserman and Lafferty 2006 ; Lafferty and Wasserman 2008], a variation of Rodeo was proposed in [Liu et al. 2007] for the density estimation. Our method we called CDRodeo (for Conditional Density Rodeo) addresses the more general problem of conditional density estimation.
Like Rodeo (which means Regularisation Of Derivative Expectation Operator), the CDRodeo algorithm generates an iterative path of decreasing bandwidths, based on tests on the partial derivatives of the estimator with respect to the components of the bandwidth. Note that the greediness of the procedure leans on the selection of this path of bandwidths, which enables us to address high dimensional problems of functional inference.
Let us be more precise: we take a kernel K of class $\mathcal{C}^1$ and consider the statistics Z_{hj} for $h \in (\mathbb{R}_+^*)^d$ and $j = 1 : d$, defined by:

$$Z_{hj} := \frac{\partial}{\partial h_j}\hat{f}_h(w).$$

Z_{hj} is easily computable, since it can be expressed by:

$$Z_{hj} = \frac{-1}{nh_j^2}\sum_{i=1}^{n} \frac{1}{\tilde{f}_X(X_i)} J\left(\tfrac{w_j - W_{ij}}{h_j}\right) \prod_{k \neq j} h_k^{-1} K\left(\frac{w_k - W_{ik}}{h_k}\right), \tag{2}$$

where $J : \mathbb{R} \to \mathbb{R}$ is the function defined by:

$$t \mapsto K(t) + tK'(t). \tag{3}$$

The details of the CDRodeo procedure are described in **Algorithm 1** and can be summed up in one sentence: for a well-chosen threshold λ_{hj} (specified in Section 3.3), the algorithm performs at each iteration the test $|Z_{hj}| > \lambda_{hj}$ to determine if the component j of the current bandwidth must be shrunk or not. It can be interpreted by the following principle: the bandwidth of a kernel estimator quantifies within which distance of the point of interest w and at which degree an observation W_i helps in the estimation. Heuristically, the larger the variation of f is, the smaller the bandwidth is required for an accurate estimation. The statistics $Z_{hj} = \frac{\partial}{\partial h_j}\hat{f}_h(w)$ are used as a proxy of $\frac{\partial}{\partial w_j} f(w)$ to quantify the variation of f in the direction w_j. Note in particular that since the partial derivatives vanish for irrelevant components, this bandwidth selection leads to an implicit variable selection, and thus to avoid the curse of dimensionality under sparsity assumptions.

Algorithm 1 CDRODEO algorithm

1. *Input:* the point of interest w, the data W, $\beta \in (0,1)$ the bandwidth decreasing factor, $h_0 > 0$ the bandwidth initialization value, a parameter $a > 1$.

2. *Initialization:*

   (a) Initialize the bandwidth: *for* $j = 1 : d$, $h_j \leftarrow h_0$.

   (b) Activate all the variables: $\mathcal{A} \leftarrow \{1, \ldots, d\}$.

3. *While* ($\mathcal{A} \neq \emptyset$) & ($\prod_{k=1}^{d} h_k \geq \frac{\log n}{n}$):

   for all $j \in \mathcal{A}$*:*

   (a) Update Z_{hj} and λ_{hj}.

   (b) *If* $|Z_{hj}| \geq \lambda_{hj}$: update $h_j \leftarrow \beta h_j$.
   else: remove j from $\mathcal{A}$.

4. *Output:* h (and $\hat{f}_h(w)$).

3 Theoretical results

This section gathers the theoretical results of our method.

3.1 Assumptions

We consider K a compactly supported kernel. For any bandwidth $h \in (\mathbb{R}_+^*)^d$, we define the neighbourhood $\mathcal{U}_h(u)$ of $u \in \mathbb{R}^{d'}$ (typically, $u = x$ or w, and $d' = d_1$ or d) as follows:

$$\mathcal{U}_h(u) := \left\{ u' \in \mathbb{R}^{d'} : \forall j = 1 : d', u'_j = u_j - h_j z_j, \text{with } z \in (\text{supp}(K))^{d'} \right\}.$$

Then we denote the CDRODEO initial bandwidth $h^{(0)} = \left(\frac{1}{\log n}, \ldots, \frac{1}{\log n}\right)$ and for short, $\mathcal{U}_n(u) := \mathcal{U}_{h^{(0)}}(u)$. We also introduce the notation $\|\cdot\|_{\infty, \mathcal{U}}$ for the supremum norm over a set $\mathcal{U}$.

The following first assumption ensures a certain amount of observations in the neighbourhood of our point of interest w.

Assumption 1 (f_X bounded away of 0). *We assume* $\delta := \inf_{u \in \mathcal{U}_n(x)} f_X(u) > 0$.

Note that if the neighbourhood $\mathcal{U}_n(x)$ does not contain any observation X_i, the estimation of the conditional distribution of Y given the event $X = x$ is obviously intractable.

The second assumption specifies the notions of "sparse function" and "relevant component", under which the curse of high dimensionality can be avoided.

Assumption 2 (Sparsity condition). *There exists a subset* $\mathcal{R} \in \{1, \ldots, d\}$ *such that for any fixed* $\{z_j\}_{j \in \mathcal{R}}$*, the function* $\{z_k\}_{k \in \mathcal{R}^c} \mapsto f(z_1, \ldots, z_d)$ *is constant on* $\mathcal{U}_n(w)$*.*

In other words, if we denote r the cardinal of $\mathcal{R}$, Assumption 2 means that f locally depends on only r of its d variables. We call *relevant* any component in $\mathcal{R}$. The notion of relevant component depends on the point where f is estimated. For example, a component w_j which behaves as $\mathbb{1}_{[0,1]}(w_j)$ in the conditional density is only relevant in the neighbourhood of 0 and 1. Note that this local property addresses a broader class of functions, which extends the application field of Theorem 2 and improves the convergence rate of the method.

Finally, the conditional density is required to be regular enough.

Assumption 3 (Regularity of f). *There exists a known integer p such that f is of class $\mathcal{C}^p$ on $\mathcal{U}_n(w)$ and such that $\partial_j^p f(w) \neq 0$ for all $j \in \mathcal{R}$.*

3.2 Conditions on the estimator of f_X

Given the definition of the estimator (1), we need an estimator $\tilde{f}_X$ of f_X.

If f_X is known. We take $\tilde{f}_X \equiv f_X$. This case is not completely obvious. In particular, it tackles the case of *un*conditional density estimation, if we set by convention $d_1 = 0$ and $f_X \equiv 1$.

If f_X is unknown. We need an estimator $\tilde{f}_X$ which satisfies the following two conditions:

(i) a positive lower bound: $\widetilde{\delta}_X := \inf_{u \in \mathcal{U}_n(x)} \tilde{f}_X(u) > 0$

(ii) a concentration inequality in local sup norm: there exists a constant $M_X > 0$ such that:

$$\mathbb{P}\left(\sup_{u \in \mathcal{U}_n(x)} \left| \frac{f_X(u) - \tilde{f}_X(u)}{\tilde{f}_X(u)} \right| > M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}} \right) \leq \exp(-(\log n)^{\frac{5}{4}}).$$

The following proposition proves these conditions are feasible. Furthermore, the provided estimator of f_X (see the proof in Section 5.3.1) is easily implementable and does not need any optimisation.

Proposition 1. *Given a sample $\widetilde{X}$ with same distribution as X and of size $n_X = n^c$ with $c > 1$, if f_X is of class $\mathcal{C}^{p'}$ with $p' \geq \frac{d_1}{2(c-1)}$, there exists an estimator $\tilde{f}_X$ which satisfies (i) and (ii).*

3.3 CDRodeo parameters choice.

Kernel K. We choose the kernel function $K : \mathbb{R} \to \mathbb{R}$ of class $\mathcal{C}^1$, with compact support and of order p, *i.e.*: for $\ell = 1, \ldots, p-1$, $\int_{\mathbb{R}} t^\ell K(t)dt = 0$, and $\int_{\mathbb{R}} t^p K(t)dt \neq 0$.
Note that considering a compactly supported kernel is fundamental for the local approach. In particular, it relaxes the assumptions by restricting them to a neighbourhood of w.
Taking a kernel of order p is usual for the control of the bias of the estimator.

Parameter β. Let $\beta \in (0, 1)$ be the decreasing factor of the bandwidth. The larger β, the more accurate the procedure, but the longer the computational time. From the theoretical point of view, it remains of little importance, as it only affects the constant terms. In practice, we set it close to 1.

Bandwidth initialization. We recall that we set $h_0 := \frac{1}{\log n}$ (and the initial bandwidth as $\left(\frac{1}{\log n}, \ldots, \frac{1}{\log n}\right)$).

Threshold $\lambda_{h,j}$. For any bandwidth $h \in (\mathbb{R}_+^*)^d$ and for $j = 1 : d$, we set the threshold as follows:

$$\lambda_{hj} := \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{n h_j^2 \prod_{k=1}^d h_k}}, \tag{4}$$

with $\mathsf{C}_\lambda := 4\|J\|_2 \|K\|_2^{d-1}$ (where J is defined in (3)) and $a > 1$. The expression is obtained by using concentration inequalities on Z_{hj}. For the proof, the parameter a has to be tuned such that:

$$(\log n)^{a-1} > \frac{\|f\|_{\infty, \mathcal{U}_n(w)}}{\delta}, \tag{5}$$

which is satisfied for n large enough. The influence of this parameter is discussed in the next section, once the theoretical results are stated.

Hereafter, unless otherwise specified, the parameters are chosen as described in this section.

3.4 Mains results

Let us denote $\hat{h}$ the bandwidth selected by CDRodeo. In Theorem 2, we introduce a set $\mathcal{H}_{\text{hp}}$ of bandwidths which contains $\hat{h}$ with high probability, which leads to an upper bound of the pointwise estimation error with high probability. In Corollary 3, we deduce the convergence rate of CDRodeo from Theorem 2.
More precisely, in Theorem 2, we determine lower and upper bounds (with high probability) for the stopping iteration of each bandwidth component. We set:

$$\tau_n := \frac{1}{(2p+r)\log\frac{1}{\beta}} \log\left(\frac{n}{\mathsf{C}_\tau (\log n)^{2p+d+a}}\right), \tag{6}$$

and

$$T_n := \tau_n + \frac{\log\left(\mathsf{C}_T^{-1}\right)}{(2p+1)\log\frac{1}{\beta}}, \tag{7}$$

where

$$\mathsf{C}_\tau := \left(\frac{4(p-1)!\mathsf{C}_\lambda}{\left(\min\limits_{j\in\mathcal{R}} \partial_j^p f(w)\right)\int_{\mathbb{R}} t^p K(t)dt}\right)^2, \; \mathsf{C}_T := \left(\frac{\min\limits_{j\in\mathcal{R}} |\partial_j^p f(w)|}{24\max\limits_{j\in\mathcal{R}} |\partial_j^p f(w)|}\right)^2.$$

Then we define the set of bandwidths $\mathcal{H}_{\text{hp}}$ by:

$$\mathcal{H}_{\text{hp}} := \left\{h \in \mathbb{R}_+^d : h_j = \frac{\beta^{\theta_j}}{\log n}, \text{ with } \theta_j \in \{\lfloor \tau_n \rfloor + 1, \dots, \lfloor T_n \rfloor\} \text{ if } j \in \mathcal{R}, \text{ else } \theta_j = 0\right\}.$$

Theorem 2. *Assume that $\tilde{f}_X$ satisfies Conditions (i) and (ii) of section 3.2 and Assumptions 1 to 3 are satisfied. Then, the bandwidth $\hat{h}$ selected by* CDRodeo *belongs to $\mathcal{H}_{hp}$ with high probability. More precisely, for any $q > 0$ and for n large enough:*

$$\mathbb{P}\left(\hat{h} \in \mathcal{H}_{hp}\right) \geq 1 - n^{-q}. \tag{8}$$

Moreover, with probability larger than $1 - 2n^{-q}$, the CDRodeo *estimator $\hat{f}_{\hat{h}}(w)$ verifies:*

$$\left|\hat{f}_{\hat{h}}(w) - f(w)\right| \leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} \tag{9}$$

with

$$\mathsf{C} := 2r\mathsf{C}_\tau^{\frac{p}{2p+r}} \int_{t\in\mathbb{R}} |\frac{t^p}{p!}K(t)|dt \times \max_{k\in\mathcal{R}} \|\partial_k^p f\|_{\infty, \mathcal{U}_n(w)} + 4\|K\|_2^d \|f\|_{\infty, \mathcal{U}_n(w)}^{\frac{1}{2}} \delta^{-\frac{1}{2}} \mathsf{C}_T^{\frac{-r}{2(2p+1)}} \mathsf{C}_\tau^{\frac{-r}{2(2p+r)}}.$$

Corollary 3. *Under the assumptions of Theorem 2, for any $q \geq 1$:*

$$\left(\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^q\right]\right)^{1/q} \leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} + o\left(n^{-1}\right).$$

Corollary 3 presents a generalization of the previous works on Rodeo [Lafferty and Wasserman 2008] and [Liu et al. 2007] whose results are restricted to the regularity $p = 2$ and to simpler problems, namely regression and density estimation.
We compare the convergence rate of CDRodeo with the optimal minimax rate. In particular, our benchmark is the pointwise minimax rate, which is of order $\mathcal{O}\left(n^{-\frac{p}{2p+d}}\right)$, for the problem of p-regular d-dimensional density estimation, obtained by [Donoho and Low 1992].
Without sparsity structure ($r = d$), CDRodeo achieves the optimal minimax rate, up to a logarithmic factor. The exponent of this factor depends on the parameter a. For the proofs, we need $a > 1$ in order to satisfy (5), but if an upper bound (or a pre-estimator) of $\frac{\|f\|_{\infty, \mathcal{U}_n(w)}}{\delta}$ were known, we could obtain the similar result with $a = 1$ and a modified constant term. Note that the logarithmic factor is a small price to pay for a computationally-tractable procedure for high-dimensional functional inference, in particular see section 3.5 for the computational gain of our procedure.

Under sparsity assumptions, we avoid the curse of high dimensionality and our procedure achieves the desired rate $n^{-\frac{p}{2p+r}}$ (up to a logarithmic term), which is optimal if the relevant components were known. Note that some additional logarithmic factors could be unavoidable due to the unknown sparsity structure, which needs to be estimated. Identifying the exact order of the logarithm term in the optimal minimax rate for the sparse case remains an open challenging question.

3.5 Complexity

We now discuss the complexity of CDRodeo without taking into account the pre-computation cost of $\tilde{f}_X$ at the points X_i, $i = 1 : n$ (used for computing the Z_{hj}), but a fast procedure for $\tilde{f}_X$ is required, to avoid losing CDRodeo computational advantages.

For CDRodeo, the main cost lies in the computation of the Z_{hj}'s along the path of bandwidths.

The condition $\prod_{k=1}^{d} h_k \geq \frac{\log n}{n}$ restricts to at most $\log_{\beta^{-1}} n$ updates of the bandwidth across all components, leading to a worst-case complexity of order $\mathcal{O}(d.n \log n)$.

But as shown in Theorem 2, with high probability, $\hat{h} \in \mathcal{H}_{\text{hp}}$, in which only the relevant components are active after the first iteration. In first iteration, the $Z_{h^{(0)}j}$'s computation costs $\mathcal{O}(d.n)$ operations, while the product kernel enables us to compute the Z_{hj}'s in following iteration with only $\mathcal{O}(r.n)$ operations, which leads to the complexity $\mathcal{O}(d.n + r.n \log n)$.

In order to grasp the advantage of CDRodeo greediness, we compare its complexity with optimization over an exhaustive bandwidth grid with $\log n$ values for each component of the bandwidth (which is often the case in others methods: Cross validation, Lepski methods...): for each bandwidth of $(\log n)^d$-sized grid, the computation of a statistic from the $d.n$-sized dataset needs at least $\mathcal{O}(d.n)$ operation, which leads to a complexity of order $\mathcal{O}(d.n(\log_{\beta^{-1}} n)^d)$. Using the parameters used in the simulated example in section 4 ($n = 2.10^5$, $d = 5$, $r = 3$, $\beta = 0.95$), the ratio of complexities is $\dfrac{d.n(\log n)^d}{r.n \log n} \approx 5.10^9$, and even without sparsity structure: $\dfrac{d.n(\log n)^d}{d.n \log n} \approx 3.10^9$. It means that CDRodeo run is a billion times faster on this data set.

4 Simulations

In this section, we test the practical performances of our method. In particular, we study CDRodeo on a 5-dimensional example. The major purpose of this section is to assess if the numerical performances of our procedure. Let us describe the example. We set $d_1 = 4$ and $d_2 = 1$ and simulate an i.i.d sample $\{(X_i, Y_i)\}_{i=1}^n$ with the following distribution: for any $i = 1, \ldots, n$:

- the first component X_{i1} of X_i follows a uniform distribution on $[-1, 1]$,
- the other components X_{ij}, $j = 2 : 4$, are independent standard normal and are independent of X_{i1},
- Y_i is independent of X_{i1}, X_{i3} and X_{i4} and the conditional distribution of Y_i given X_{i2} is exponential with survival parameter X_{i2}^2.

The estimated conditional density function is then defined by:

$$f : (x, y) \mapsto \mathbb{1}_{[-1,1]}(x_1) \frac{1}{x_2^2} e^{-\frac{y}{x_2^2}}.$$

This example enables us to test several criteria: sparsity detection, behaviour when fonctions are not continuous, bimodality estimation, robustness when f_X takes small values.

In the following simulations, if not stated explicitly otherwise, Rodeo is run with sample size $n = 200,000$, product Gaussian kernel, initial bandwidth value $h_0 = 0.4$, bandwidth decreasing factor $\beta = 0.95$ and parameter $a = 1.1$ and $\tilde{f}_X \equiv f_X$.

Figure 1 illustrates CDRodeo bandwidth selection. In which, the boxplots of each selected bandwidth component are built from 200 runs of CDRodeo at the point $w = (0, 1, 0, 0, 1)$. This figure reflects the specificity of

CDRodeo to capture the relevance degree of each component, and one could compare it with variable selection (as done in [Lafferty and Wasserman 2008]). The components x_3 and x_4 are irrelevant and for this point of interest, the components x_2 and y are clearly relevant while the component x_1 is barely relevant as f is constant in the direction x_1 in near neighbourhood of $x_1 = 0$. As expected, the irrelevant h_3 and h_4 are mostly deactivated at the first iteration, while the relevant h_2 and h_5 are systematically shrunk. The relevance degree of x_1 is also well detected as the values of h_1 are smaller than h_0, but significantly larger than h_2 and h_5.

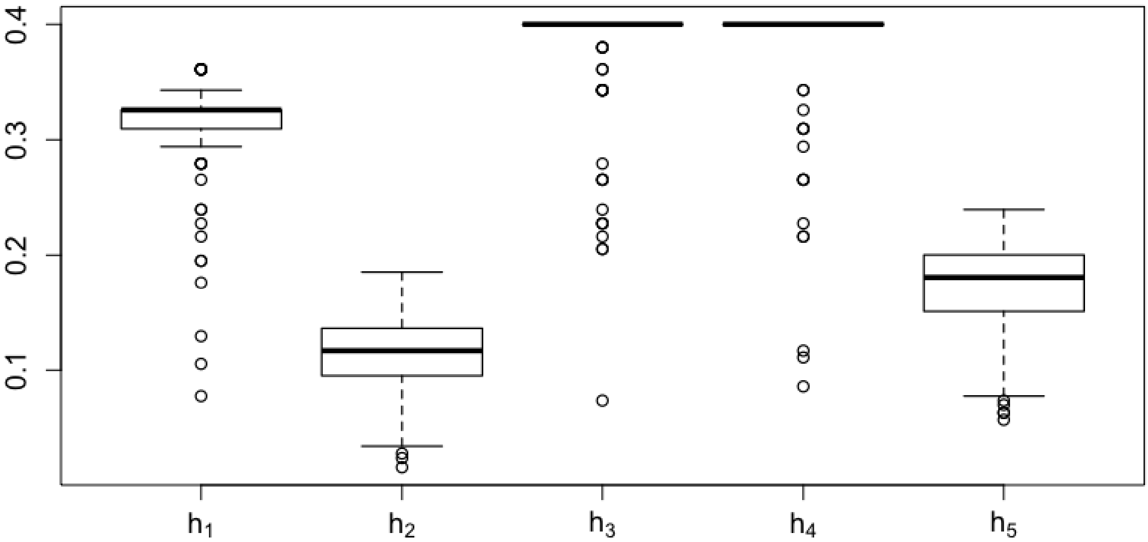

Figure 1: Boxplots of each component of 200 CDRodeo selected bandwidths at the point $w = (0, 1, 0, 0, 1)$.

Figure 2 gives CDRodeo estimation of f from one n-sample. The function f is well estimated. In particular, irrevance, jumps and bi-modality are features which are well detected by our method. As expected, main estimation errors are made on points of discontinuity for x_1 and y or at the boundaries for x_2, x_3 and x_4. Note that the f_X values are particularly small at the boundaries of the plots in function of x, leading to lack of observations for the estimation. Note however that null value for f_X does not deteriorate the estimation (cf top left plot), since the estimate of f vanishes automatically when there is no observation near the point of interest.

Running time. The simulations are implemented in R on a Macintosh laptop with a $3, 1$ GHz Intel Core i7 processor. In the Figure Figure 1, the 200 runs of CDRodeo take 2952.735 seconds (around 50 minutes), or 14.8 seconds per run.

5 Proofs

We first give the outlines of the proofs in Section 5.1. To facilitate the lecture of the proof, we have divided the proofs of the main results (Proposition 1, Theorem 2 and Corollary 3) into intermediate results which are stated in Section 5.2 and proved in Section 5.4. The proof of the main results are in Section 5.3.

5.1 Outlines of the proofs

We first prove Proposition 1 by constructing an estimator of f_X with the wanted properties. In this proof, we use some usual properties of a kernel density estimator (control of the bias, concentration inequality), which are gathered in Lemma 4.

Theorem 2 states two results: the bandwidth selection (8) and the estimation error of the procedure (9). For the proof of the bandwidth selection (8), Proposition 8 mades explicit the highly probable behaviour of CDRodeo along a run, and thus the final selected bandwidth. In particular, the proof leans on an analysis of Z_{hj}, which is made in two steps. We first consider $\bar{Z}_{hj}$, a simpler version of Z_{hj} in which we substitute the estimator of

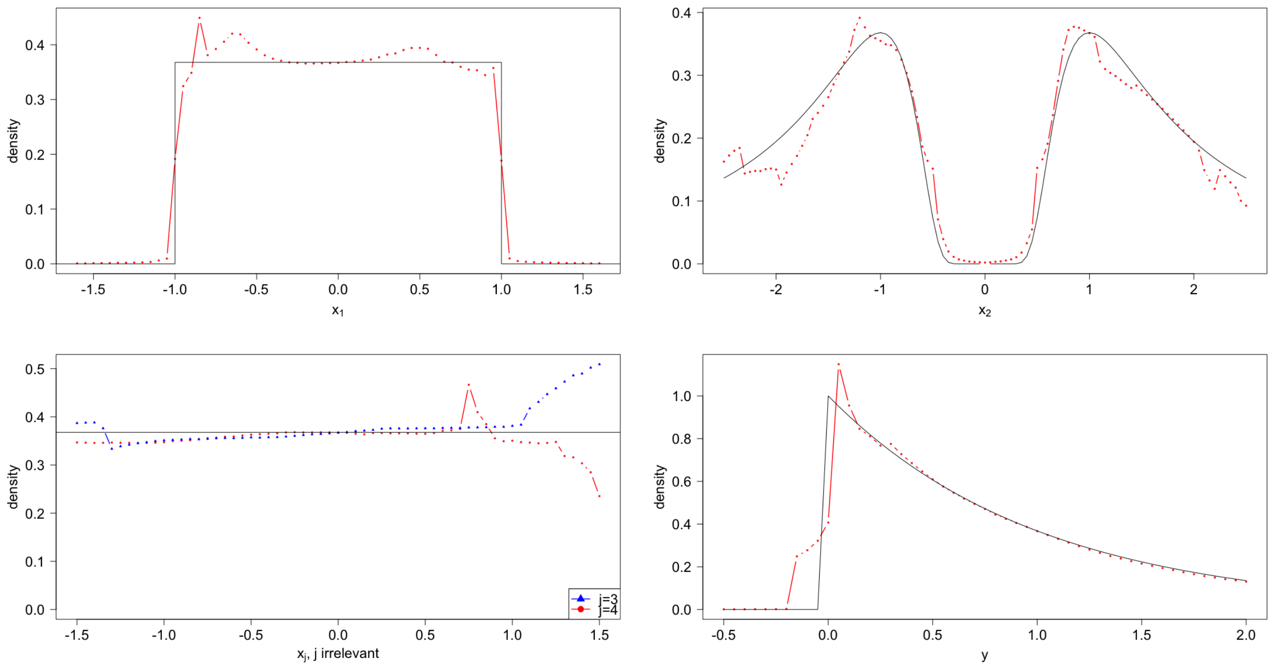

Figure 2: CDRodeo estimator (red or blue dashed lines) VS the true density (black solid line) in function of each component, the others component being fixed following $(x, y) = (0, 1, 0, 0, 1)$.

f_X by f_X itself, and we detail its behaviour in Lemma 6. Then we control the difference $Z_{hj} - \bar{Z}_{hj}$ (see in 1. of Lemma 7) to ensure $\bar{Z}_{hj}$ behaves like Z_{hj}.
To control the estimation error of the procedure (9), we similarly analyse $\hat{f}_h(w)$ in two parts: in Lemma 5, we describe the behaviour of $\bar{f}_h(w)$, the simpler version of $\hat{f}_h(w)$ in which we substitute the estimator of f_X by f_X itself, and in 2. of Lemma 7, we bound the difference $f_h - \bar{f}_h(w)$. Then the bandwidth selection (8) leads to the upper bound with high probability of the estimation error of $\hat{f}_h(w)$ (9).
Finally, we obtain the expected error of $\hat{f}_h(w)$ stated in Corollary 3 by controlling the error on the residual event.

5.2 Intermediate results

For any bandwidth $h_X \in \mathbb{R}_+^*$, we define the kernel density estimator $\tilde{f}_X^K$ by: for any $u \in \mathbb{R}^{d_1}$,

$$\tilde{f}_X^K(u) := \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \prod_{j=1}^{d_1} K_X\left(\frac{u_j - \widetilde{X}_{ij}}{h_X}\right), \tag{10}$$

where $K_X : \mathbb{R} \to \mathbb{R}$ a kernel which is compactly supported, of class $\mathcal{C}^1$, of order $p_X \geq \frac{d_1}{2(c-1)}$, where we recall that $c > 1$ is defined by $n_X = n^c$.
We also introduce the neighbourhood

$$\mathcal{U}_n'(x) := \{u' = u - h_X z : u \in \mathcal{U}_n(x), z \in \text{supp}(K_X)\}. \tag{11}$$

Lemma 4 ($\tilde{f}_X^K$ behaviour). *We assume f_X is $\mathcal{C}^{p'}$ on $\mathcal{U}_n'(x)$ with $p' \leq p_X$, then for any bandwidth $h_X \in \mathbb{R}_+^*$,*

1. *if we denote $C_{bias_X} := \frac{\|K_X\|_1^{d_1 - 1} \|.^{p'} K_X(\cdot)\|_1}{p'!} d_1 \max_{k=1:d_1} \|\partial_k^{p'} f_X\|_{\infty, \mathcal{U}_n'(x)}$, then*

$$\left\| \mathbb{E}\left[\tilde{f}_X^K\right] - f_X \right\|_{\infty, \mathcal{U}_n(x)} \leq C_{bias_X} h_X^{p'}.$$

2. *If the condition*

$$\textit{Cond}_X(h_X): \quad h_X^{d_1} \geq \frac{4\|K_X\|_\infty^{2d_1}}{9\|K_X\|_2^{2d_1}\|f_X\|_{\infty, \mathcal{U}_n'(x)}} \frac{(\log n)^{\frac{3}{2}}}{n_X}$$

is satisfied, then for $\lambda_X := \sqrt{\frac{4\|K_X\|_2^{2d_1}\|f_X\|_{\infty,\,\mathcal{U}_n'(x)}}{h_X^{d_1} n_X}(\log n)^{\frac{3}{2}}}$ *and for any* $u \in \mathcal{U}_n(x)$*:*

$$\mathbb{P}\left(\left|\tilde{f}_X^K(u) - \mathbb{E}\left[\tilde{f}_X^K(u)\right]\right| > \lambda_X\right) \le 2\exp\left(-(\log n)^{\frac{3}{2}}\right).$$

Lemma 5 ($\bar{f}_h(w)$ behaviour)**.** *For any bandwidth* $h \in (0, h_0]^d$*, and any* $i = 1 : n$*, let us denote* $\bar{f}_{hi}(w) := \frac{K_h(w-W_i)}{f_X(X_i)}$*. Then, if* K *is chosen as in section 3.3, under Assumptions 1 to 3,*

1. *Let* $\mathsf{C}_{\bar{\mathsf{E}}} := \|f\|_{\infty,\,\mathcal{U}_n(w)}\|K\|_1^d$*. Then*

$$\left|\mathbb{E}\left[\bar{f}_{h1}(w)\right]\right| \le \mathbb{E}\left[\left|\bar{f}_{h1}(w)\right|\right] \le \mathsf{C}_{\bar{\mathsf{E}}}.$$

   Besides, if we denote $\overline{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$ *the bias of* $\bar{f}_h(w) := \frac{1}{n}\sum\limits_{i=1}^{n}\bar{f}_{hi}(w)$*, then:*

$$|\overline{B}_h| \le \mathsf{C}_{\overline{\mathsf{bias}}}\sum_{k\in\mathcal{R}} h_k^p,$$

   with $\mathsf{C}_{\overline{\mathsf{bias}}} := \frac{2|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!}\max\limits_{k\in\mathcal{R}}|\partial_k^p f(w)|$*.*

2. *Let* $\overline{\mathcal{B}}_h := \{|\bar{f}_h(w) - \mathbb{E}[\bar{f}_h(w)]| \le \sigma_h\}$*, where* $\sigma_h := \mathsf{C}_\sigma\sqrt{\frac{(\log n)^a}{n\prod\limits_{k=1}^{d} h_k}}$ *with* $\mathsf{C}_\sigma = \frac{2\|K\|_2^d\|f\|_{\infty,\,\mathcal{U}_n(w)}^{\frac{1}{2}}}{\delta^{\frac{1}{2}}}$*. If Cond(h):* $\prod\limits_{k=1}^{d} h_k \ge \frac{4\|K\|_\infty^{2d}}{9\delta^2\mathsf{C}_\sigma^2}\frac{(\log n)^a}{n}$ *is satisfied, then:*

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \le 2e^{-(\log n)^a}$$

3. *Let* $\mathcal{B}_{|\bar{f}|h} := \{|\frac{1}{n}\sum\limits_{i=1}^{n}|\bar{f}_{hi}(w)| - \mathbb{E}[|\bar{f}_h(w)|]| \le \mathsf{C}_{\bar{\mathsf{E}}}\}$*. Then*

$$\mathbb{P}\left(\mathcal{B}_{|\bar{f}|h}^c\right) \le 2e^{-\mathsf{C}_{\gamma|f|}n\prod_{k=1}^d h_k},$$

with $\mathsf{C}_{\gamma|f|} := \min\left(\frac{\mathsf{C}_{\bar{\mathsf{E}}}^2}{\mathsf{C}_\sigma^2}; \frac{3\delta\mathsf{C}_{\bar{\mathsf{E}}}}{4\|K\|_\infty^d}\right)$*.*

Lemma 6 ($\bar{Z}_{hj}$ behaviour)**.** *For any* $j \in \{1,\ldots,d\}$ *and any bandwidth* $h \in (0, h_0]^d$*, we define* $\bar{Z}_{hij} := \frac{1}{f_X(X_i)}\frac{\partial}{\partial h_j}\left(\prod\limits_{k=1}^{d} h_k^{-1}K\left(\frac{w_k - W_{ik}}{h_k}\right)\right)$*, and* $\bar{Z}_{hj} := \frac{1}{n}\sum_{i=1}^n \bar{Z}_{hij}$*. If* K *is chosen as in section 3.3, and under Assumptions 1 to 3,*

1. *Under Assumptions 1 to 3, for* $j \notin \mathcal{R}$*:*

$$\mathbb{E}\left[\bar{Z}_{hj}\right] = 0.$$

   whereas, for $j \in \mathcal{R}$*, for* n *large enough,*

$$\frac{1}{2}\mathsf{C}_{E\bar{Z},j}h_j^{p-1} \le \left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| \le \frac{3}{2}\mathsf{C}_{E\bar{Z},j}h_j^{p-1}, \tag{12}$$

   where $\mathsf{C}_{E\bar{Z},j} := \left|\frac{\int_{\mathbb{R}} t^p K(t)dt}{(p-1)!}\partial_j^p f(w)\right|$*.*
   Besides, let $\mathsf{C}_{E|\bar{Z}|} := \|f\|_{\infty,\,\mathcal{U}_n(w)}\|J\|_1\|K\|_1^{d-1}$*. Then :*

$$\mathbb{E}\left[|\bar{Z}_{h1j}|\right] \le \mathsf{C}_{E|\bar{Z}|}h_j^{-1}. \tag{13}$$

2. *Let* $\mathcal{B}_{\bar{Z},hj} := \{|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]| \le \frac{1}{2}\lambda_{hj}\}$*. Under Assumptions 1 to 3, if the bandwidth satisfies:*

Cond$_{\bar{Z}}(h)$: $\prod_{k=1}^{d} h_k \geq$ *cond*$_{\bar{Z}} \frac{(\log n)^a}{n}$, *with cond*$_{\bar{Z}} := \frac{4\|J\|_\infty^2 \|K\|_\infty^{2(d-1)}}{3^2 \|f\|_{\infty,\, \mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}}$,

then: $\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) \leq 2e^{-\gamma_{Z,n}}$, *with* $\gamma_{Z,n} := \frac{\delta}{\|f\|_{\infty,\, \mathcal{U}_n(w)}} (\log n)^a$.

3. *Let* $\mathcal{B}_{|\bar{Z}|,hj} := \{ | \frac{1}{n} \sum_{i=1}^{n} |\bar{Z}_{hij}| - \mathbb{E}\left[|\bar{Z}_{h1j}|\right] | \leq \mathsf{C}_{E|\bar{Z}|} h_j^{-1} \}$. *Then, under Assumptions 1 to 3:*

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,hj}\right) \leq 2e^{-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^{d} h_k},$$

with $\mathsf{C}_{\gamma|\bar{Z}|} := \min\left(\frac{\delta \mathsf{C}^2_{E|\bar{Z}|}}{4\|f\|_{\infty,\, \mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}} ; \frac{3\delta \mathsf{C}_{E|\bar{Z}|}}{4\|K\|_\infty^{d-1} \|J\|_\infty} \right)$.

Lemma 7. *For any* $h \in (0, h_0]^d$ *and any component* $j = 1 : d$, *we denote* $\Delta_{Z,hj} := Z_{hj} - \bar{Z}_{hj}$ *and* $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$. *Under Assumptions 1 to 3, if the conditions on* $\tilde{f}_X$ *are satisfied (see section 3.2), then,*

1. *for* $\mathsf{C}_{\mathsf{M}\Delta Z} := \frac{2\mathsf{C}_{E|\bar{Z}|} M_X}{\mathsf{C}_\lambda}$*:*

$$\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n} |\Delta_{Z,hj}| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}} \lambda_{hj}$$

2. *for* $\mathsf{C}_{\mathsf{M}\Delta} := \frac{2\mathsf{C}_{\bar{\mathsf{E}}} M_X}{\mathsf{C}_\sigma}$*:*

$$\mathbb{1}_{\widetilde{A}_n \cap \mathcal{B}_{|\bar{f}|h}} |\Delta_h| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}} \sigma_h.$$

We introduce the notation $h^{(t)}$, $t \in \mathbb{N}$, the state of the bandwidth at iteration t if $\hat{h} = h$. In particular for a fixed $t \in \{0, \ldots, \lfloor \tau_n \rfloor\}$, $h^{(t)}$ is identical for any $h \in \mathcal{H}_{\mathsf{hp}}$. Then we consider the event:

$$\mathcal{E}_Z := \widetilde{A}_n \cap \bigcap_{j \notin \mathcal{R}} \left\{ \mathcal{B}_{\bar{Z},h^{(0)}j} \cap \mathcal{B}_{|\bar{Z}|,h^{(0)}j} \right\} \cap \bigcap_{j \in \mathcal{R}} \left[\bigcap_{h \in \mathcal{H}_{\mathsf{hp}}} \left\{ \mathcal{B}_{\bar{Z},hj} \cap \mathcal{B}_{|\bar{Z}|,hj} \right\} \cap \bigcap_{t=0}^{\lfloor \tau_n \rfloor} \left\{ \mathcal{B}_{\bar{Z},h^{(t)}j} \cap \mathcal{B}_{|\bar{Z}|,h^{(t)}j} \right\} \right],$$

where we denote $\widetilde{A}_n := \left\{ \sup_{u \in \mathcal{U}_n(x)} \left| \frac{f_X(u) - \tilde{f}_X(u)}{\tilde{f}_X(u)} \right| \leq M_X \frac{(\log n_X)^b}{n_X^\alpha} \right\}$.

Proposition 8 (CDRodeo behaviour)**.** *Under Assumptions 1 to 3, on* $\mathcal{E}_Z$, $\hat{h} \in \mathcal{H}_{hp}$.
In other words, when $\mathcal{E}_Z$ *happens:*

1. *non relevant components are deactivated during the iteration* 0*;*

2. *at the end of the iteration* $\lfloor \tau_n \rfloor$*, the active components are exactly the relevant ones;*

3. CDRodeo *stops at last at the iteration* $\lfloor T_n \rfloor$.

Moreover, for any $q > 0$*:*

$$\mathbb{P}\left(\mathcal{E}_Z^c\right) = o(n^{-q}).$$

The following lemma give a technical result to canonically obtain an upper bound of the bias of a kernel estimator. Let us denote $\cdot$ the multiplication terms by terms of two vectors.

Lemma 9. *Let* $u \in \mathbb{R}^{d'}$ *and a bandwidth* $h \in \left(\mathbb{R}_+^*\right)^{d'}$. *For* $j = 1 : d'$, *let* $K : \mathbb{R} \to \mathbb{R}$ *be a continuous function with compact support and with at least* $p - 1$ *zero moments, ie: for* $l = 1 : (p-1)$,

$$\int_{\mathbb{R}} K(t) t^l dt = 0.$$

Let $f_0 : \mathbb{R}^{d'} \to \mathbb{R}$ a function of class $\mathcal{C}^p$ on $\mathcal{U}_h(u) := \left\{u' \in \mathbb{R}^{d'} : \forall j = 1 : d', u'_j = u_j - h_j z_j, \text{with } z_j \in supp(K_j)\right\}$. Then:

$$\int_{\mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} h_j^{-1} K\left(\frac{u_j - u'_j}{h_j}\right)\right) f_0(u')du' - f_0(u) \int_{\mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j)\right) dz = \sum_{k=1}^{d'} (\mathsf{I}_k + \mathsf{II}_k), \tag{14}$$

where

$$\mathsf{I}_k := \int_{z \in \mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j)\right) \rho_k dz,$$

with the notations $\rho_k := \rho_k(z,h,u) = (-h_k z_k)^p \int_{0 \le t_p \le \dots \le t_1 \le 1} \left(\partial_k^p f_0(\overline{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f_0(\overline{z}_{k-1})\right) dt_{1:p}$, and $\overline{z}_{k-1} := u - \sum_{j=1}^{k-1} h_j z_j e_j$ (where $\{e_j\}_{j=1}^{d'}$ is the canonical basis of $\mathbb{R}^{d'}$), and

$$\mathsf{II}_k := (-h_k)^p \int_{t \in \mathbb{R}} \frac{t^p}{p!} K(t)dt \int_{z_{-k} \in \mathbb{R}^{d'-1}} \partial_k^p f_0(\overline{z}_{k-1}) \left(\prod_{j \neq k} K(z_j)\right) dz_{-k}.$$

Finally, we recall (without proof) the classical Bernstein's Inequality and Taylor's theorem with integral remainder.

Lemma 10 (Bernstein's inequality). *Let $U_1, \dots, U_n$ be independent random variables almost surely uniformly bounded by a positive constant $c > 0$ and such that for $i = 1, \dots, n$, $\mathbb{E}[U_i^2] \le v$. Then for any $\lambda > 0$,*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n U_i - \mathbb{E}[U_i]\right| \ge \lambda\right) \le 2\exp\left(-\min\left(\frac{n\lambda^2}{4v}, \frac{3n\lambda}{4c}\right)\right).$$

Note that this version is a simple consequence of Birgé and Massart (p.366 of [Birgé and Massart 1998]).

Lemma 11 (Taylor's theorem). *Let $g : [0,1] \to \mathbb{R}$ be a function of class $\mathcal{C}^q$. Then we have:*

$$g(1) - g(0) = \sum_{l=1}^q \frac{g^{(l)}(0)}{l!} + \int_{t_1=0}^1 \int_{t_2=0}^{t_1} \cdots \int_{t_q=0}^{t_{q-1}} (g^{(q)}(t_q) - g^{(q)}(0)) dt_q dt_{q-1} \dots dt_1$$

5.3 Proofs of Proposition 1, Theorem 2 and Corollary 3

5.3.1 Proof of Proposition 1

We construct $\tilde{f}_X$ in two steps: we first construct an estimator $\tilde{f}_X^K$ which satisfies

$$\mathbb{P}\left(\|f_X - \tilde{f}_X^K\|_{\infty, \mathcal{U}_n(x)} > M_X \frac{(\log n)^{\frac{3}{4}}}{n^{\frac{1}{2}}}\right) \le \exp(-(\log n)^{\frac{5}{4}}), \tag{15}$$

then we show that if we set $\tilde{f}_X \equiv \tilde{f}_X^K \vee (\log n)^{-\frac{1}{4}}$, $\tilde{f}_X$ satisfies Conditions (i) and (ii) for n large enough.

We take $\tilde{f}_X^K$ as the kernel density estimator defined in(10), with a kernel $K_X : \mathbb{R} \to \mathbb{R}$ that is compactly supported, of class $\mathcal{C}^1$, of order $p_X \ge \frac{d_1}{2(c-1)}$. and a bandwidth $h_X \in \mathbb{R}_+^*$ specified later.Let us control the bias $\left\|\mathbb{E}\left[\tilde{f}_X^K\right] - f_X\right\|_{\infty, \mathcal{U}_n(x)}$. We define $p'_X = \min(p', p_X + 1)$. In particular, f_X is of class $\mathcal{C}^{p'_X}$ and K_X has $p'_X - 1$ zero moments.
Therefore we can apply Lemma 4 :

$$\left\|\mathbb{E}\left[\tilde{f}_X^K\right] - f_X\right\|_{\infty, \mathcal{U}_n(x)} \le \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X},$$

where $\mathsf{C}'_{\mathsf{bias}_X} := \frac{\|K_X\|_1^{d_1-1}\|.^{p'_X}K_X(\cdot)\|_1}{p'_X!} d_1 \max_{k=1:d_1} \|\partial_k^{p'_X} f_X\|_{\infty,\, \mathcal{U}'_n(x)}$.
Therefore, since

$$\begin{aligned}
\left\|\tilde{f}_X^K - f_X\right\|_{\infty,\, \mathcal{U}_n(x)} &\leq \left\|\tilde{f}_X^K - \mathbb{E}\left[\tilde{f}_X^K\right]\right\|_{\infty,\, \mathcal{U}_n(x)} + \left\|\mathbb{E}\left[\tilde{f}_X^K\right] - f_X\right\|_{\infty,\, \mathcal{U}_n(x)} \\
&\leq \left\|\tilde{f}_X^K - \mathbb{E}\left[\tilde{f}_X^K\right]\right\|_{\infty,\, \mathcal{U}_n(x)} + \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X},
\end{aligned}$$

we have for any threshold λ:

$$\mathbb{P}\left(\left\|\tilde{f}_X^K - f_X\right\|_{\infty,\, \mathcal{U}_n(x)} \geq \lambda\right) \leq \mathbb{P}\left(\left\|\tilde{f}_X^K - \mathbb{E}\left[\tilde{f}_X^K\right]\right\|_{\infty,\, \mathcal{U}_n(x)} \geq \lambda - \mathsf{C}'_{\mathsf{bias}_X} h_X^{p'_X}\right). \tag{16}$$

Therefore, we have reduced the problem to a local concentration inequality of $\tilde{f}_X^K$ in sup norm. In order to move from a supremum on $\mathcal{U}_n(x)$ to a maximum on a finite set of elements of $\mathcal{U}_n(x)$, let us construct a ϵ-net of $\mathcal{U}_n(x)$. We denote $A > 0$ such that:

$$\mathsf{supp}(K_X) \cup \mathsf{supp}(K) \subset \left[-\tfrac{A}{2}, \tfrac{A}{2}\right].$$

We set $N(\epsilon)$ the smallest integer such that $\epsilon N(\epsilon) \geq \frac{A}{\log n}$, *i.e.*:

$$N(\epsilon) := \left\lceil \frac{A}{\epsilon \log n} \right\rceil,$$

then we introduce the notation $u_{(l)} \in \mathcal{U}_n(x)$, for a multi-index $l \in (1 : N(\epsilon))^{d_1}$ defined, such that the j^{th} component of $u_{(l)}$ is:

$$u_{(l)j} := x_j - \frac{A}{2\log n} + (2l_j - 1)\frac{\epsilon}{2}.$$

Then $\{u_{(l)} : l \in (1 : N(\epsilon))^{d_1}\}$ is a ϵ-net of $\mathcal{U}_n(x)$, in the meaning that for any $u \in \mathcal{U}_n(x)$, there exists $l \in \{1, \ldots, N(\epsilon)\}^{d_1}$ such that $\|u - u_{(l)}\|_\infty := \max_{k=1:d_1} |u_k - u_{(l)k}| \leq \epsilon$.
Therefore to obtain the desired concentration inequality, we only need to obtain the concentration inequality for each point of $\{u_{(l)} : l \in (1 : N(\epsilon))^{d_1}\}$ and to control the following supremum

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\epsilon))^{d_1}} \left|\tilde{f}_X^K(u) - \mathbb{E}\left[\tilde{f}_X^K(u)\right] - \tilde{f}_X^K(u_{(l)}) + \mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right|.$$

For this purpose, we obtain (from Taylor's Inequality): for any $u, v \in \mathbb{R}^{d_1}$,

$$\left|\prod_{k=1}^{d_1} K_X(u_k) - \prod_{k=1}^{d_1} K_X(v_k)\right| \leq d_1 \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \|u - v\|_\infty.$$

Therefore, for any $u, v \in \mathcal{U}_n(x)$:

$$\begin{aligned}
\left|\tilde{f}_X^K(u) - \tilde{f}_X^K(v)\right| &\leq \frac{1}{n_X . h_X^{d_1}} \sum_{i=1}^{n_X} \left|\prod_{k=1}^{d_1} K_X\left(\tfrac{u_k - \widetilde{X}_{ik}}{h_X}\right) - \prod_{k=1}^{d_1} K_X\left(\tfrac{u_k - \widetilde{X}_{ik}}{h_X}\right)\right| \\
&\leq \frac{d_1}{h_X^{d_1+1}} \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \|u - v\|_\infty.
\end{aligned}$$

Since $\{u_{(l)} : l \in (1 : N(\epsilon))^{d_1}\}$ is a ϵ-net of $\mathcal{U}_n(x)$:

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\epsilon))^{d_1}} \left|\tilde{f}_X^K(u) - \tilde{f}_X^K(u_{(l)})\right| \leq \frac{d_1}{h_X^{d_1+1}} \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \epsilon.$$

Thus:

$$\sup_{u \in \mathcal{U}_n(x)} \min_{l \in (1:N(\epsilon))^{d_1}} \left|\mathbb{E}\left[\tilde{f}_X^K(u)\right] - \mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right| \leq d_1 \|K'_X\|_\infty \|K_X\|_\infty^{d_1-1} \frac{\epsilon}{h_X^{d_1+1}}.$$

And so:

$$\sup_{u\in\,\mathcal{U}_n(x)}\ \min_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{f}_X^K(u)-\mathbb{E}\left[\tilde{f}_X^K(u)\right]-\tilde{f}_X^K(u_{(l)})+\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right|\leq 2d_1\|K_X'\|_\infty\|K_X\|_\infty^{d_1-1}\frac{\epsilon}{h_X^{d_1+1}}.$$

We denote $\mathsf{C}_{\mathsf{diff}} := 2d_1\|K_X'\|_\infty\|K_X\|_\infty^{d_1-1}$. Then:

$$\begin{aligned}\left\|\tilde{f}_X^K-\mathbb{E}\left[\tilde{f}_X^K\right]\right\|_{\infty,\,\mathcal{U}_n(x)} &\leq \max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{f}_X^K(u_{(l)})-\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right| \\ &\quad+\sup_{u\in\,\mathcal{U}_n(x)}\ \min_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{f}_X^K(u)-\mathbb{E}\left[\tilde{f}_X^K(u)\right]-\tilde{f}_X^K(u_{(l)})+\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right| \\ &\leq \max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{f}_X^K(u_{(l)})-\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right|+\mathsf{C}_{\mathsf{diff}}\frac{\epsilon}{h_X^{d_1+1}}.\end{aligned}$$

Then the inequality (16) becomes: for any threshold λ,

$$\begin{aligned}\mathbb{P}\left(\left\|\tilde{f}_X^K-f_X\right\|_{\infty,\,\mathcal{U}_n(x)}\geq\lambda\right) &\leq \mathbb{P}\left(\left\|\tilde{f}_X^K-\mathbb{E}\left[\tilde{f}_X^K\right]\right\|_{\infty,\,\mathcal{U}_n(x)}\geq\lambda-\mathsf{C}'_{\mathsf{bias}_X}h_X^{p'_X}\right) \\ &\leq \mathbb{P}\left(\max_{l\in(1:N(\epsilon))^{d_1}}\left|\tilde{f}_X^K(u_{(l)})-\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right|\geq\lambda-\mathsf{C}'_{\mathsf{bias}_X}h_X^{p'_X}-\mathsf{C}_{\mathsf{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right) \\ &\leq N(\epsilon)^{d_1}\max_{l\in(1:N(\epsilon))^{d_1}}\mathbb{P}\left(\left|\tilde{f}_X^K(u_{(l)})-\mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right|\geq\lambda-\mathsf{C}'_{\mathsf{bias}_X}h_X^{p'_X}-\mathsf{C}_{\mathsf{diff}}\frac{\epsilon}{h_X^{d_1+1}}\right)\end{aligned} \tag{17}$$

We want to apply 2. of Lemma 4. Therefore we fix the following settings:

- $h_X := n_X^{-\frac{c-1}{c.d_1}}$
- $\lambda := 2\lambda_X$, where λ_X is the threshold in 2. of Lemma 4;
- $\epsilon := h_X^{1+\frac{d_1}{2}}n_X^{-\frac{1}{2}}$.

For short, we denote $\mathsf{C}_{\lambda X} := 2\|K_X\|_2^{d_1}\|f_X\|_{\infty,\,\mathcal{U}'_n(x)}^{\frac{1}{2}}$, so:

$$\lambda_X=\sqrt{\frac{4\|K_X\|_2^{2d_1}\|f_X\|_{\infty,\,\mathcal{U}'_n(x)}}{h_X^{d_1}n_X}(\log n)^{\frac{3}{2}}}=\mathsf{C}_{\lambda X}(\log n)^{\frac{3}{4}}h_X^{-\frac{d_1}{2}}n_X^{-\frac{1}{2}}=\mathsf{C}_{\lambda X}(\log n)^{\frac{3}{4}}n_X^{-\frac{1}{2c}}.$$

In particular, since we take $p_X\geq\frac{d_1}{2(c-1)}$ and we assume $p'\geq\frac{d_1}{2(c-1)}$, then $p'_X=\min(p',p_X)\geq\frac{d_1}{2(c-1)}$. Hence we obtain for n large enough:

$$\begin{aligned}\mathsf{C}'_{\mathsf{bias}_X}h_X^{p'_X} &= \mathsf{C}'_{\mathsf{bias}_X}n_X^{-\frac{p'_X(c-1)}{c.d_1}} \\ &\leq \mathsf{C}'_{\mathsf{bias}_X}n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X=\frac{\mathsf{C}_{\lambda X}}{2}(\log n)^{\frac{3}{4}}n_X^{-\frac{1}{2c}}.\end{aligned}$$

and also, since $c>1$:

$$\begin{aligned}\mathsf{C}_{\mathsf{diff}}\frac{\epsilon}{h_X^{d_1+1}} &= \mathsf{C}_{\mathsf{diff}}h_X^{-\frac{d_1}{2}}n_X^{-\frac{1}{2}}=\mathsf{C}_{\mathsf{diff}}\,n_X^{-\frac{1}{2c}} \\ &\leq \frac{1}{2}\lambda_X=\frac{\mathsf{C}_{\lambda X}}{2}(\log n)^{\frac{3}{4}}n_X^{-\frac{1}{2c}}.\end{aligned}$$

Thus:

$$\lambda-\mathsf{C}'_{\mathsf{bias}_X}h_X^{p'_X}-\mathsf{C}_{\mathsf{diff}}\frac{\epsilon}{h_X^{d_1+1}}\geq\lambda_X,$$

and the inequality (17) becomes:

$$\mathbb{P}\left(\left\|\tilde{f}_X^K - f_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq N(\epsilon)^{d_1} \max_{l\in(1:N(\epsilon))^{d_1}} \mathbb{P}\left(\left|\tilde{f}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right| \geq \lambda_X\right) \tag{18}$$

We verify that $\text{Cond}_X(h_X)$ is satisfied for n large enough:

$$\begin{aligned} h_X^{d_1} &= n_X^{-\frac{c-1}{c}} \\ &\geq \frac{4\|K_X\|_\infty^{2d_1}}{9\|K_X\|_2^{2d_1}\|f_X\|_{\infty,\,\mathcal{U}_n'(x)}} \frac{(\log n)^{\frac{3}{2}}}{n_X}. \end{aligned}$$

Then we can apply 2. of Lemma 4,

$$\mathbb{P}\left(\left|\tilde{f}_X^K(u_{(l)}) - \mathbb{E}\left[\tilde{f}_X^K(u_{(l)})\right]\right| > \lambda_X\right) \leq 2\exp\left(-(\log n)^{\frac{3}{2}}\right).$$

Thus the inequality (18) becomes:

$$\mathbb{P}\left(\left\|\tilde{f}_X^K - f_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq 2N(\epsilon)^{d_1} \exp\left(-(\log n)^{\frac{3}{2}}\right). \tag{19}$$

Let us control $2N(\epsilon)^{d_1}$:

$$\begin{aligned} 2N(\epsilon)^{d_1} &= 2\left\lceil \frac{A}{\epsilon \log n} \right\rceil^{d_1} \\ &= 2\left\lceil \frac{A}{h_X^{1+\frac{d_1}{2}} n_X^{-\frac{1}{2}} \log n} \right\rceil^{d_1} \\ &= o\left(n_X^{d_1+1}\right) \end{aligned}$$

Then for n large enough:

$$2N(\epsilon)^{d_1} \exp\left(-(\log n)^{\frac{3}{2}}\right) \leq \exp\left(-(\log n)^{\frac{5}{4}}\right).$$

Therefore:

$$\mathbb{P}\left(\left\|\tilde{f}_X^K - f_X\right\|_{\infty,\,\mathcal{U}_n(x)} \geq \lambda\right) \leq \exp\left(-(\log n)^{\frac{5}{4}}\right)$$

Since $\lambda = 2\mathrm{C}_{\lambda X}(\log n)^{\frac{3}{4}} n_X^{-\frac{1}{2c}}$, we have obtained the desired concentration inequality (15) with $M_X = 2\mathrm{C}_{\lambda X}$.

Now we consider $\tilde{f}_X \equiv \tilde{f}_X^K \vee (\log n)^{-\frac{1}{4}}$. By construction, $\tilde{f}_X$ satisfies Condition (i). Let us show it also satisfies Condition (ii), *i.e.*:

$$\mathbb{P}\left(\sup_{u\in\,\mathcal{U}_n(x)} \left|\frac{f_X(u) - \tilde{f}_X(u)}{\tilde{f}_X(u)}\right| > M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) \leq C_X \exp(-(\log n)^{\frac{5}{4}}).$$

We write:

$$\begin{aligned} \mathbb{P}\left(\sup_{u\in\,\mathcal{U}_n(x)} \left|\frac{f_X(u) - \tilde{f}_X(u)}{\tilde{f}_X(u)}\right| > M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) &= \mathbb{P}\left(\exists u \in \mathcal{U}_n(x), \left|f_X(u) - \tilde{f}_X(u)\right| > \tilde{f}_X(u) M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) \\ &\leq \mathbb{P}\left(\exists u \in \mathcal{U}_n(x), \left|f_X(u) - \tilde{f}_X(u)\right| > (\log n)^{-\frac{1}{4}} M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) \\ &\leq \mathbb{P}\left(\left\|f_X(u) - \tilde{f}_X(u)\right\|_{\infty,\,\mathcal{U}_n(x)} > M_X \frac{(\log n)^{\frac{d}{2}-\frac{1}{4}}}{n^{\frac{1}{2}}}\right) \end{aligned}$$

Since $d = d_1 + d_2 \geq 2$, $\frac{d}{2} - \frac{1}{4} \geq \frac{3}{4}$, we obtain from the previously proved concentration inequality (15):

$$\mathbb{P}\left(\sup_{u \in \mathcal{U}_n(x)} \left|\frac{f_X(u) - \tilde{f}_X(u)}{\tilde{f}_X(u)}\right| > M_X \frac{(\log n)^{\frac{d}{2}}}{n^{\frac{1}{2}}}\right) \leq \mathbb{P}\left(\left\|\tilde{f}_X^K - f_X\right\|_{\infty, \mathcal{U}_n(x)} \geq M_X \frac{(\log n)^{\frac{3}{4}}}{n^{\frac{1}{2}}}\right)$$
$$\leq \exp\left(-(\log n)^{\frac{5}{4}}\right).$$

5.3.2 Proof of Theorem 2

We introduce $\mathcal{E}_f := \bigcap_{h \in \mathcal{H}_{\text{hp}}} \left(\overline{\mathcal{B}}_h \cap \mathcal{B}_{|\bar{f}|h}\right)$ and denote $\mathcal{E} := \mathcal{E}_Z \cap \mathcal{E}_f$. On $\mathcal{E}$, $\hat{h}$ belongs to $\mathcal{H}_{\text{hp}}$ (cf Proposition 8). Thus:

$$\mathbb{1}_{\mathcal{E}}\left(\hat{f}_{\hat{h}}(w) - f(w)\right) = \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\text{hp}}} \mathbb{1}_{\hat{h}=h}\left(\hat{f}_h(w) - f(w)\right). \tag{20}$$

For any $h \in \mathcal{H}_{\text{hp}}$, we denote $\Delta_h := \hat{f}_h(w) - \bar{f}_h(w)$ and $\overline{B}_h := \mathbb{E}\left[\bar{f}_h(w)\right] - f(w)$, and we decompose the loss as follows:

$$\left|\hat{f}_h(w) - f(w)\right| \leq |\Delta_h| + \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| + \left|\overline{B}_h\right|. \tag{21}$$

Using Lemma 7, since $\mathcal{E} \subset \widetilde{A}_n \cap \mathcal{B}_{|\bar{f}|h}$:

$$\mathbb{1}_{\mathcal{E}} |\Delta_h| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}} \sigma_h. \tag{22}$$

Moreover, by Lemma 5, since $\mathcal{E} \subset \widetilde{A}_n \cap \overline{\mathcal{B}}_h$:

$$\begin{aligned}
\left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| &\leq \sigma_h && (23)\\
&= \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n \prod_{k=1}^{d} h_k}}\\
&\leq \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n h_0^d \beta^{r(T_n - \tau_n) + r\tau_n}}}\\
&= \mathsf{C}_\sigma \mathsf{C}_T^{\frac{-r}{2(2p+1)}} \mathsf{C}_\tau^{\frac{-r}{2(2p+r)}} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}}. && (24)
\end{aligned}$$

And, also:

$$|\overline{B}_h| \leq \mathsf{C}_{\overline{\text{bias}}} \sum_{k \in \mathcal{R}} h_k^p \leq r\mathsf{C}_{\overline{\text{bias}}} \beta^{p\tau_n} h_0^p = r\mathsf{C}_{\overline{\text{bias}}} \mathsf{C}_\tau^{\frac{p}{2p+r}} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}}. \tag{25}$$

To conclude,

$$\begin{aligned}
\mathbb{1}_{\mathcal{E}} |\hat{f}_{\hat{h}}(w) - f(w)| &\leq \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\text{hp}}} \mathbb{1}_{\hat{h}=h} \left|\hat{f}_h(w) - f(w)\right|, \text{ by (20)}\\
&\leq \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\text{hp}}} \mathbb{1}_{\hat{h}=h} \left(|\Delta_h| + \left|\bar{f}_h(w) - \mathbb{E}\left[\bar{f}_h(w)\right]\right| + \left|\overline{B}_h\right|\right), \text{ by (21)}\\
&\leq \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\text{hp}}} \mathbb{1}_{\hat{h}=h} \left[\left(1 + \frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}}\right) \sigma_h + \left|\overline{B}_h\right|\right], \text{ by (22) and (23)}\\
&\leq \mathbb{1}_{\mathcal{E}} \sum_{h \in \mathcal{H}_{\text{hp}}} \mathbb{1}_{\hat{h}=h} \mathsf{C} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}}, \text{ by (24) and (25)}\\
&= \mathbb{1}_{\mathcal{E}} \mathsf{C} (\log n)^{\frac{p(a+d-r)}{2p+r}} n^{-\frac{p}{2p+r}},
\end{aligned}$$

with for n large enough (*ie*: $\frac{\mathsf{C}_{\mathsf{M}\Delta}}{(\log n)^{\frac{a}{2}}} \leq 1$),

$$\mathsf{C} := r\mathsf{C}_{\overline{\text{bias}}}\mathsf{C}_{\tau}^{\frac{p}{2p+r}} + 2\mathsf{C}_{\sigma}\mathsf{C}_{T}^{\frac{-r}{2(2p+1)}}\mathsf{C}_{\tau}^{\frac{-r}{2(2p+r)}}.$$

It remains to give an upper bound on $\mathbb{P}\left(\mathcal{E}^c\right)$. For any $q > 0$:

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}^c\right) &\leq \mathbb{P}\left(\mathcal{E}_Z^c\right) + \mathbb{P}\left(\mathcal{E}_f^c\right)\\ &\leq o(n^{-q}) + \sum_{h\in\mathcal{H}_{\text{hp}}}\left(\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) + \mathbb{P}\left(\mathcal{B}_{|\bar{f}|h}^c\right)\right), \text{ using Proposition 8}\\ &\leq o(n^{-q}) + \sum_{h\in\mathcal{H}_{\text{hp}}}\left(2e^{-(\log n)^a} + 2e^{-\mathsf{C}_{\gamma|f|}n\prod_{k=1}^d h_k}\right),\end{aligned}$$

using Lemma 5, since for any $h \in \mathcal{H}_{\text{hp}}$, $\text{Cond}(h)$ is satisfied. Moreover:

$$n\prod_{k=1}^d h_k \geq n\beta^{rT_n}h_0^d = \mathsf{C}_T^{\frac{r}{2p+1}}\mathsf{C}_{\tau}^{\frac{r}{2p+r}}(\log n)^{\frac{ra-2p(d-r)}{(2p+r)}}n^{\frac{2p}{2p+r}} \geq \frac{(\log n)^a}{\mathsf{C}_{\gamma|f|}},$$

for n large enough. Hence:

$$\mathbb{P}\left(\mathcal{E}^c\right) \leq o(n^{-q}) + |\mathcal{H}_{\text{hp}}|4e^{-(\log n)^a} = o(n^{-q}),$$

for n large enough, since $|\mathcal{H}_{\text{hp}}| = (\lceil T_n\rceil - \lfloor \tau_n\rfloor)^r == \left(\frac{1}{(2p+1)(\log(\frac{1}{\beta}))}\log(\frac{C_T}{C_\tau}) + 1\right)^r$ is finite.

5.3.3 Proof of Corollary 3

We consider the event $\mathcal{E} = \left\{\left|\hat{f}_{\hat{h}}(w) - f(w)\right| \leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)}n^{-\frac{p}{2p+r}}\right\}$ for which we proved in Theorem 2:

$$\mathbb{P}\left(\mathcal{E}^c\right) = o(n^{-A}),$$

for any $A > 0$. For short, we denote $R_h := \left|\hat{f}_h(w) - f(w)\right|$ for any bandwidth $h \in (\mathbb{R}_+^*)^d$. Then we decompose $R_{\hat{h}}$ as follows:

$$R_{\hat{h}} = \mathbb{1}_{\mathcal{E}}R_{\hat{h}} + \mathbb{1}_{\mathcal{E}^c}R_{\hat{h}}.$$

By definition of $\mathcal{E}$, we immediately obtain:

$$\mathbb{1}_{\mathcal{E}}R_{\hat{h}} \leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)}n^{-\frac{p}{2p+r}} \tag{26}$$

For the second term, we first bound $\hat{f}_{\hat{h}}(w)$ a.s. In CDRodeo procedure, the loop stops when the current bandwidth becomes too small: $\prod_{k=1}^d h_k < \frac{(\log n)}{n}$. So the final bandwidth $\hat{h}$ satisfies:

$$\prod_{k=1}^d \hat{h}_k \geq \frac{(\log n)}{\beta^d n}.$$

Since $\hat{f}_{\hat{h}}(w) = \frac{1}{n}\sum_{i=1}^n \frac{1}{f_X(X_i)}\left(\prod_{k=1}^d \hat{h}_k^{-1}K\left(\frac{w_j - W_{ij}}{\hat{h}_j}\right)\right)$,

$$\left|\hat{f}_{\hat{h}}(w)\right| \leq \frac{\beta^d\|K\|_\infty^d}{\delta}\frac{n}{\log n}$$

Hence:

$$R_{\hat{h}} \leq f(w) + \frac{\beta^d\|K\|_\infty^d}{\delta}\frac{n}{(\log n)}$$

Therefore, for any $q > 0$, using $(a+b)^q \leq 2^{q-1}(a^q + b^q)$:

$$\mathbb{E}\left[\mathbb{1}_{\mathcal{E}^c}(R_{\hat{h}})^q\right] \leq P(\mathcal{E}^c)2^{q-1}\left(f(w)^q + \|K\|_\infty^q \frac{n^q}{(\log n)^q}\right)$$
$$= o\left(n^{-A'}\right), \tag{27}$$

for any $A' > 0$ (since $P(\mathcal{E}^c) = o(n^{-A'+q})$).
We conclude by combining (26) and (27):

$$\left(\mathbb{E}\left[\left|\hat{f}_{\hat{h}}(w) - f(w)\right|^q\right]\right)^{1/q} \leq \mathsf{C}(\log n)^{\frac{p}{2p+r}(d-r+a)} n^{-\frac{p}{2p+r}} + o\left(n^{-1}\right). \tag{28}$$

5.4 Proof of Proposition 8 and the lemmas

5.4.1 Proof of Proposition 8

First, note that the final state of the bandwidth determines exactly at which iteration each component has been deactivated: for a fixed bandwidth $h \in (\mathbb{R}_+^*)^d$, if $\hat{h} = h$, we denote $\{\theta_k\}_{k=1}^d$ such as for $k = 1:d$, $h_k = h_0\beta^{\theta_k}$. In particular, θ_k is the iteration of deactivation of the component k.
We introduce the notation $h^{(t)}$, $t \in \mathbb{N}$, the state of the bandwidth at iteration t if $\hat{h} = h$. It implies that $h^{(t)}$ is exactly defined by: $h_k^{(t)} = \beta^{\theta_k \wedge t} h_0$ for $k = 1:d$.
Notice that for a fixed $t \in \{0, \dots, \lfloor \tau_n \rfloor\}$, $h^{(t)}$ is identical for any $h \in \mathcal{H}_{\mathsf{hp}}$: by definition of $\mathcal{H}_{\mathsf{hp}}$, $h_j^{(t)} = h_0\beta^t$ if $j \in \mathcal{R}$, else $h_j^{(t)} = h_0$.
We recall the definition

$$\mathcal{E}_Z := \widetilde{A}_n \cap \bigcap_{j \notin \mathcal{R}} \left\{\mathcal{B}_{\bar{Z},h^{(0)}j} \cap \mathcal{B}_{|\bar{Z}|,h^{(0)}j}\right\} \cap \bigcap_{j \in \mathcal{R}} \left[\bigcap_{h \in \mathcal{H}_{\mathsf{hp}}} \left\{\mathcal{B}_{\bar{Z},hj} \cap \mathcal{B}_{|\bar{Z}|,hj}\right\} \cap \bigcap_{t=0}^{\lfloor \tau_n \rfloor} \left\{\mathcal{B}_{\bar{Z},h^{(t)}j} \cap \mathcal{B}_{|\bar{Z}|,h^{(t)}j}\right\}\right].$$

For any component j and any bandwidth h, we decompose Z_{hj} as follows:

$$\mathbb{1}_{\mathcal{E}_Z} Z_{hj} = \mathbb{1}_{\mathcal{E}_Z} \bar{Z}_{hj} + \mathbb{1}_{\mathcal{E}_Z} \Delta_{Z,hj}$$
$$= \mathbb{1}_{\mathcal{E}_Z} \mathbb{E}\left[\bar{Z}_{hj}\right] + \mathbb{1}_{\mathcal{E}_Z}(\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]) + \mathbb{1}_{\mathcal{E}_Z} \Delta_{Z,hj} \tag{29}$$

1. Let us fix $j \notin \mathcal{R}$ and $h = h^{(0)} = (h_0, \dots, h_0)$. Using 2. of Lemma 6, $\mathbb{E}\left[\bar{Z}_{hj}\right] = 0$. Therefore:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \leq \mathbb{1}_{\mathcal{E}_Z}\left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| + \mathbb{1}_{\mathcal{E}_Z}\left|\Delta_{Z,hj}\right|$$
$$\leq \frac{1}{2}\lambda_{hj} + \mathbb{1}_{\mathcal{E}_Z}\left|\Delta_{Z,hj}\right|,$$

   using 2. of Lemma 6, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj}$. Now using 1. of Lemma 7, since $\mathcal{E}_Z \subset \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$, we obtain:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \leq \frac{1}{2}\lambda_{hj} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}.$$

   Then for n large enough (*ie*: $(\log n)^{\frac{a}{2}} > 2\mathsf{C}_{\mathsf{M}\Delta Z}$), $\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| < \lambda_{hj}$. In other words, when $\mathcal{E}_Z$ happens, all irrelevant components deactivate at the iteration 0.

2. Let us show that $\mathcal{E}_Z$ implies that the relevant components remain active until iteration $\lfloor \tau_n \rfloor + 1$.
   It suffices to prove $|Z_{h^{(t)}j}| > \lambda_{h^{(t)}j}$, for any $j \in \mathcal{R}$ and any bandwidth $h^{(t)}$, $t = 0 : \lfloor \tau_n \rfloor$. (Indeed, by induction: $(h_0, \dots, h_0) = h^{(0)}$, and since the irrelevant components deactivate at the iteration 0, if the current bandwidth at the iteration t is $h^{(t)}$, then the fact that all the relevant components remain active for this bandwidth implies that the bandwidth at iteration $t+1$ is $h^{(t+1)}$).
   Let us fix $j \in \mathcal{R}$, $t = 0 : \lfloor \tau_n \rfloor$ and we denote $h = h^{(t)}$. Using the decomposition (29), we obtain the following lower bound:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\left(\left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| - \left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| - |\Delta_{Z,hj}|\right).$$

   Then, combining:

- $\left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| \geq \frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1}$ (cf 1. of Lemma 6),
- $\left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| \leq \frac{1}{2}\lambda_{hj}$, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj}$ (cf 2. of Lemma 6),
- $|\Delta_{Z,hj}| \leq \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}$, since $\mathcal{E}_Z \subset \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$ (cf 1. of Lemma 7),

we obtain:

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\left(\frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1} - \frac{1}{2}\lambda_{hj} - \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}\right).$$

Now let us show: $\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \geq \mathbb{1}_{\mathcal{E}_Z}\lambda_{hj}$.
First, if n is large enough (ie $(\log n)^{\frac{a}{2}} \geq 2\mathsf{C}_{\mathsf{M}\Delta Z}$), then

$$\frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj} \leq \frac{1}{2}\lambda_{hj}.$$

Then it suffices to prove:

$$\frac{\mathsf{C}_{E\bar{Z},j}}{2} h_j^{p-1} \geq 2\lambda_{hj},$$

i.e.:

$$h_j^{2p}\prod_{k=1}^{d} h_k \geq \frac{4^2\mathsf{C}_\lambda^2}{\mathsf{C}_{E\bar{Z},j}^2}\frac{(\log n)^a}{n}.$$

It is ensured for $t \leq \tau_n$, by definition of τ_n in (6):

$$h_j^{2p}\prod_{k=1}^{d} h_k = \frac{\beta^{t(2p+r)}}{(\log n)^{2p+d}} \geq \frac{\beta^{\tau_n(2p+r)}}{(\log n)^{2p+d}} = \frac{4^2\mathsf{C}_\lambda^2}{\min_{k\in\mathcal{R}}\mathsf{C}_{E\bar{Z},k}^2}\frac{(\log n)^a}{n} \geq \frac{4^2\mathsf{C}_\lambda^2}{\mathsf{C}_{E\bar{Z},j}^2}\frac{(\log n)^a}{n}. \tag{30}$$

Therefore, on $\mathcal{E}_Z$, the component j remains active until the iteration $\lfloor\tau_n\rfloor$.

3. Let us now prove that on $\mathcal{E}_Z$, each relevant component j deactivates at last at iteration $\lceil T_n\rceil$. In particular, by definition of $\mathcal{H}_{\mathsf{hp}}$, $\hat{h}$ belongs to $\mathcal{H}_{\mathsf{hp}}$ on $\mathcal{E}_Z$.
Assume $\mathcal{E}_Z$ happens.
We fix $j \in \mathcal{R}$. It suffices to prove that if j is still active at iteration $\lceil T_n\rceil$, then on $\mathcal{E}_Z$ happens, it deactivates at the end of this iteration. We assume j is still active and we denote h the state of the bandwidth at iteration $\lceil T_n\rceil$.

   By the first point, for any $k \notin \mathcal{R}$, $h_k = h_0$.
Given the second point, each relevant component k was still active at the beginning of the iteration $\lfloor\tau_n\rfloor + 1$, *ie*: for any $k \in \mathcal{R}$, $h_k \leq \beta^{\lfloor\tau_n\rfloor+1}h_0 \leq \beta^{\tau_n}h_0$.
Moreover, since j is still active, $h_j = \beta^{\lceil T_n\rceil}h_0$. Let us prove that: $\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| < \lambda_{hj}$. Using the decomposition (29):

$$\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| \leq \left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| + \mathbb{1}_{\mathcal{E}_Z}\left|\bar{Z}_{hj} - \mathbb{E}\left[\bar{Z}_{hj}\right]\right| + \mathbb{1}_{\mathcal{E}_Z}\left|\Delta_{Z,hj}\right|$$

   Using the points 1. and 2. of Lemma 6 and 1. Lemma 7, since $\mathcal{E}_Z \subset \mathcal{B}_{\bar{Z},hj} \cap \mathcal{B}_{|\bar{Z}|,hj} \cap \widetilde{A}_n$:

$$\begin{aligned}\mathbb{1}_{\mathcal{E}_Z}|Z_{hj}| &\leq 2\mathsf{C}_{E\bar{Z},j}h_j^{p-1} + \frac{1}{2}\lambda_{hj} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj}\\ &\leq \lambda_{hj}\left(\frac{2\mathsf{C}_{E\bar{Z},j}n^{\frac{1}{2}}h_j^p\prod_{k=1}^d h_k^{\frac{1}{2}}}{\mathsf{C}_\lambda(\log n)^{\frac{a}{2}}} + \frac{1}{2} + \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\right).\end{aligned}$$

Given the specific form of h :

$$\frac{2\mathsf{C}_{E\bar{Z},j} n^{\frac{1}{2}} h_j^p \prod_{k=1}^d h_k^{\frac{1}{2}}}{\mathsf{C}_\lambda (\log n)^{\frac{a}{2}}} \leq \frac{2\mathsf{C}_{E\bar{Z},j} n^{\frac{1}{2}} h_0^{\frac{2p+d}{2}} \beta^{\frac{(2p+1)}{2}(T_n-\tau_n)} \beta^{\frac{(2p+r)\tau_n}{2}}}{\mathsf{C}_\lambda (\log n)^{\frac{a}{2}}}$$
$$= \sqrt{\frac{4^3 \mathsf{C}^2_{E\bar{Z},j} \beta^{(2p+1)(T_n-\tau_n)}}{\min_{k\in\mathcal{R}} \mathsf{C}^2_{E\bar{Z},k}}}, \text{ by definition of } \tau_n$$
$$\leq \frac{1}{3}, \text{ by definition of } T_n.$$

Moreover, for n large enough:

$$\frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}} < \frac{1}{6}.$$

Therefore:

$$\mathbb{1}_{\mathcal{E}_Z} |Z_{hj}| < \lambda_{hj}.$$

In other words, when $\mathcal{E}_Z$ happens, any active component at iteration $\lceil T_n \rceil$ deactivates.

So we have proved that on $\mathcal{E}_Z$, $\hat{h} \in \mathcal{H}_{\mathsf{hp}}$.

It remains to show that $\mathcal{E}_Z$ holds with high probability.

$$\mathbb{P}(\mathcal{E}_Z^c) \leq \mathbb{P}\left(\widetilde{A}_n^c\right) + \sum_{k=1}^d \left\{ \mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(0)}k}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(0)}k}\right)\right\}$$
$$+ \sum_{j\in\mathcal{R}} \left[\sum_{h\in\mathcal{H}_{\mathsf{hp}}} \left(\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,hj}\right)\right) + \sum_{t=1}^{\lfloor \tau_n \rfloor} \left(\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(t)}j}\right) + \mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(t)}j}\right)\right)\right]$$

By choice of $\tilde{f}_X$:

$$\mathbb{P}\left(\widetilde{A}_n^c\right) \leq C_X e^{-(\log n)^{\frac{5}{4}}}.$$

We want to apply 2. and 3. of Lemma 6 for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $h^{(t)}$ with $t = 1 : \lfloor \tau_n \rfloor$. These bandwidths satisfy:

$$\prod_{k=1}^d h_k^{(t)} \geq \prod_{k=1}^d h_k \geq h_0^d \beta^{r\lceil T_n \rceil} \geq \mathsf{C}_T^{\frac{r}{2p+1}} \mathsf{C}_\tau^{\frac{r}{2p+r}} (\log n)^{\frac{ra-2p(d-r)}{2p+r}} n^{-\frac{r}{2p+r}},$$

which ensures that for n large enough, $\mathsf{Cond}_{\bar{Z}}(h^{(t)})$ and $\mathsf{Cond}_{\bar{Z}}(h^{(t)})$ hold for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $t = 0 : \lfloor \tau_n \rfloor$. Note in particular that $\mathcal{H}_{\mathsf{hp}} \subset \{h^{(t), t=0:\lceil T_n \rceil}\}$.
Therefore, for any component $k = 1 : d$,

$$\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},h^{(0)}k}\right) \leq 2e^{-\gamma_{Z,n}}$$

and for any $h \in \mathcal{H}_{\mathsf{hp}}$ and any $t = 0 : \lceil T_n \rceil$,

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h^{(t)}j}\right) \leq 2\exp\left(-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^d h_k\right)$$
$$\leq 2\exp\left(-\mathsf{C}_{\gamma|\bar{Z}|} \mathsf{C}_T^{\frac{r}{2p+1}} \mathsf{C}_\tau^{\frac{r}{2p+r}} (\log n)^{\frac{ra-2p(d-r)}{2p+r}} n^{\frac{2p}{2p+r}}\right)$$
$$\leq 2e^{-\gamma_{Z,n}}, \text{ for } n \text{ large enough.}$$

To conclude, note that $|\mathcal{H}_{\mathsf{hp}}| = (\lceil T_n \rceil - \lfloor \tau_n \rfloor)^r \leq (T_n - \tau_n + 2)^r = \left(\frac{\log\left(\mathsf{C}_T^{-1}\right)}{(2p+1)\log \frac{1}{\beta}} + 2 \right)^r$ is finite, so for any $q > 0$:

$$\begin{aligned}
\mathbb{P}\left(\mathcal{E}_Z^c\right) &\leq C_X e^{-(\log n)^{\frac{5}{4}}} + 2(d + r|\mathcal{H}_{\mathsf{hp}}| + r\tau_n) 2e^{-\gamma_{Z,n}} \\
&= o(n^{-q}),
\end{aligned}$$

by definition $\gamma_{Z,n} := \frac{\delta}{\|f\|_{\infty,\, \mathcal{U}_n(w)}} (\log n)^a$.

5.4.2 Proof of Lemma 4

1. We control the bias $\left\| \mathbb{E}\left[\tilde{f}_X^K\right] - f_X \right\|_{\infty,\, \mathcal{U}_n(x)}$. We write for any $u \in \mathcal{U}_n(x)$:

$$\mathbb{E}\left[\tilde{f}_X^K(u)\right] - f_X(u) = \frac{1}{h_X^{d_1}} \int_{u' \in \mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X\left(\tfrac{u_j - u'_j}{h_X}\right) \right) f_X(u') du' - f_X(u) \int_{\mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X(z_j) \right) dz$$

The kernel K_X is of order p_X and f_X is assumed of class $\mathcal{C}^{p'}$ on $\mathcal{U}'_n(x)$, with in particular $p' - 1 \leq p_X - 1$, then we can apply Lemma 9 with the settings $u = u$, $d' = d_1$, $f_0 = f_X$, $p = p' - 1$, $K = K_X$ and for $j = 1 : d'$, $h_k = h_X$. We obtain:

$$\mathbb{E}\left[\tilde{f}_X^K(u)\right] - f_X(u) = \sum_{k=1}^{d_1} (\mathsf{I}_k + \mathsf{II}_k). \tag{31}$$

with

$$\begin{aligned}
\mathsf{I}_k &:= \int_{z \in \mathbb{R}^{d_1}} \left(\prod_{k'=1}^{d_1} K_X(z_{k'}) \right) \rho_k dz, \\
\rho_k &:= \rho_k(z, h_X, u) \\
&= (-h_X z_k)^{p'-1} \int\limits_{0 \leq t_{p'-1} \leq \cdots \leq t_1 \leq 1} \left(\partial_k^{p'-1} f_X(\overline{z}_{k-1} - t_{p'-1} h_X z_k e_k) - \partial_k^{p'-1} f_X(\overline{z}_{k-1}) \right) dt_{1:(p'-1)}, \\
\mathsf{II}_k &:= (-h_X)^{p'-1} \int_{t \in \mathbb{R}} \frac{t^{p'-1}}{(p'-1)!} K_X(t) dt \int_{z_{-k} \in \mathbb{R}^{d_1 - 1}} \partial_k^{p'-1} f_X(\overline{z}_{k-1}) \left(\prod_{k' \neq k} K_X(z_{k'}) \right) dz_{-k}.
\end{aligned}$$

Let us control ρ_k. First we write:

$$\partial_k^{p'-1} f_X(\overline{z}_{k-1} - t_{p'-1} h_X z_k e_k) - \partial_k^{p'-1} f_X(\overline{z}_{k-1}) = -h_X z_k \int_{t_{p'}=0}^{1} \partial_k^{p'} f_X(\overline{z}_{k-1} - t_{p'} h_X z_k e_k) dt_{p'}.$$

Therefore:

$$\rho_k = (-h_X z_k)^{p'} \int\limits_{0 \leq t_{p'} \leq \cdots \leq t_1 \leq 1} \partial_k^{p'} f_X(\overline{z}_{k-1} - t_{p'} h_X z_k e_k) dt_{1:p'}.$$

Hence:

$$\begin{aligned}
|\rho_k| &\leq |h_X z_k|^{p'} \int\limits_{0 \leq t_{p'} \leq \cdots \leq t_1 \leq 1} \left| \partial_k^{p'} f_X(\overline{z}_{k-1} - t_{p'} h_X z_k e_k) \right| dt_{1:p'} \\
&= \frac{|z_k|^{p'}}{p'} \|\partial_k^{p'} f_X\|_{\infty,\, \mathcal{U}'_n(x)} h_X^{p'}.
\end{aligned}$$

Then:

$$\begin{aligned}
|\mathsf{I}_k| &\leq \int_{z\in\mathbb{R}^{d_1}} \left|\prod_{k'=1}^{d'} K_X(z_{k'})\right| |\rho_k|\, dz \\
&\leq \|\partial_k^{p'} f_X\|_{\infty,\,\mathcal{U}_n'(x)} h_X^{p'} \int_{z\in\mathbb{R}^{d_1}} \frac{|z_k|^{p'}}{p'} \left|\prod_{k'=1}^{d'} K_X(z_{k'})\right| dz \\
&= \frac{\|K_X\|_1^{d_1-1}\| \cdot^{p'} K_X(\cdot)\|_1}{p'} \|\partial_k^{p'} f_X\|_{\infty,\,\mathcal{U}_n'(x)} h_X^{p'}
\end{aligned} \tag{32}$$

Besides, K_X is of order p_X and $p'-1 < p_X$ and so:

$$\mathsf{II}_k := \frac{(-h_X)^{p'-1}}{(p'-1)!} \int_{t\in\mathbb{R}} t^{p'-1} K_X(t)dt \int_{z_{-k}\in\mathbb{R}^{d_1-1}} \partial_k^{p'-1} f_X(\overline{z}_{k-1}) \left(\prod_{k'\neq k} K_X(z_{k'})\right) dz_{-k} = 0.$$

Therefore the terms II_k vanish in the equation (31), and with the upper bound of I_k (32), we obtain:

$$\begin{aligned}
\left\|\mathbb{E}\left[\tilde{f}_X^K\right] - f_X\right\|_{\infty,\,\mathcal{U}_n(x)} &= \sup_{u\in\,\mathcal{U}_n(x)} \left|\mathbb{E}\left[\tilde{f}_X^K(u)\right] - f_X(u)\right| \\
&\leq \sup_{u\in\,\mathcal{U}_n(x)} \sum_{k=1}^{d_1} |\mathsf{I}_k| \\
&\leq \frac{\|K_X\|_1^{d_1-1}\| \cdot^{p'} K_X(\cdot)\|_1}{p'!} h_X^{p'} \sum_{k=1}^{d_1} \|\partial_k^{p'} f_X\|_{\infty,\,\mathcal{U}_n'(x)} \\
&= \mathsf{C}_{\mathsf{bias}_X} h_X^{p'},
\end{aligned}$$

with $\mathsf{C}_{\mathsf{bias}_X} := \frac{\|K_X\|_1^{d_1-1}\|\cdot^{p'} K_X(\cdot)\|_1}{p'!} d_1 \max_{k=1:d_1} \|\partial_k^{p'} f_X\|_{\infty,\,\mathcal{U}_n'(x)}$.

2. We apply Bernstein's inequality (see Lemma 10). We define for any $u \in \mathcal{U}_n(x)$ and any $i = 1 : n_X$:

$$\tilde{f}_{Xi}^K(u) := \frac{1}{h_X^{d_1}} \prod_{j=1}^{d_1} K_X\left(\tfrac{u_j - \widetilde{X}_{ij}}{h_X}\right).$$

Then we control $\tilde{f}_{X1}^K$ a.s.: for any $u \in \mathcal{U}_n(x)$,

$$\left|\tilde{f}_{X1}^K(u)\right| \leq \mathsf{M}_{h_X} := \|K_X\|_\infty^{d_1} h_X^{-d_1}.$$

and its variance:

$$\begin{aligned}
\mathsf{Var}\left(\tilde{f}_{X1}^K(u)\right) &\leq \mathbb{E}\left[(\tilde{f}_{X1}^K)^2\right] \\
&= h_X^{-2d_1} \int_{u'\in\mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X\left(\tfrac{u_j - u_j'}{h_X}\right)\right)^2 f_X(u')du' \\
&= h_X^{-d_1} \int_{z\in\mathbb{R}^{d_1}} \left(\prod_{j=1}^{d_1} K_X(z_j)\right)^2 f_X(u - h_X z)du' \\
&\leq \mathsf{v}_{h_X}
\end{aligned}$$

with $\mathsf{v}_{h_X} := \mathsf{C}_{\mathsf{v}X} h_X^{-d_1}$ and $\mathsf{C}_{\mathsf{v}X} := \|K_X\|_2^{2d_1} \|f_X\|_{\infty,\,\mathcal{U}_n'(x)}$.
Then we apply Lemma 10: for any $\lambda > 0$,

$$\mathbb{P}\left(\left|\tilde{f}_X^K(u) - \mathbb{E}\left[\tilde{f}_X^K(u)\right]\right| > \lambda\right) \leq 2\exp\left(-\min\left(\frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}, \frac{3n_X\lambda}{4\mathsf{M}_{h_X}}\right)\right).$$

We set $\lambda = \lambda_X := \sqrt{\frac{4\mathsf{v}_{h_X}}{n_X}(\log n)^{\frac{3}{2}}}$ such that $(\log n)^{\frac{3}{2}} = \frac{n_X\lambda^2}{4\mathsf{v}_{h_X}}$. Then we compare the rates:

$$\begin{aligned}
& \frac{n_X\lambda^2}{4\mathsf{v}_{h_X}} \leq \frac{3n_X\lambda}{4\mathsf{M}_{h_X}} \\
\iff\quad & \lambda^2 \leq \frac{3^2\mathsf{C}_{\mathsf{v}X}^2}{\|K_X\|_\infty^{2d_1}} \\
\iff\quad & h_X^{d_1} \geq \frac{4\|K_X\|_\infty^{2d_1}}{9\mathsf{C}_{\mathsf{v}X}} \frac{(\log n)^{\frac{3}{2}}}{n_X}, \\
\iff\quad & \mathsf{Cond}_X(h_X).
\end{aligned}$$

5.4.3 Proof of Lemma 5

1. We recall the notation $\cdot$ for the multiplication terms by terms of two vectors. Then:

$$\begin{aligned}
|\mathbb{E}\left[\bar{f}_{h1}(w)\right]| &\leq \mathbb{E}\left[|\bar{f}_{h1}(w)|\right] \\
&= \int_{u\in\mathbb{R}^d} \left|\prod_{k=1}^d \frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right| f(u)du \\
&= \int_{z\in\mathbb{R}^d} \left|\prod_{k=1}^d K(z_k)\right| f(w-h\cdot z)dz \\
&\leq \|f\|_{\infty,\,\mathcal{U}_n(w)}\|K\|_1^d =: \mathsf{C}_{\bar{\mathsf{E}}}
\end{aligned}$$

Now let us give an upper bound on the bias of $\bar{f}_h(w)$:

$$\overline{B}_h = \mathbb{E}\left[\bar{f}_{h1}(w)\right] - f(w) = \int_{u\in\mathbb{R}^d}\left(\prod_{k=1}^d \frac{K(h_k^{-1}(w_k-u_k))}{h_k}\right) f(u)du - f(w)\int_{\mathbb{R}^d}\prod_{k'=1}^d K(z_{k'})dz,$$

since $\int_\mathbb{R} K(t)dt = 1$. Then we apply the Lemma 9 with the settings $d' = d$, $u = w$, $h = h$, $f_0 = f$, $p = p$ and $K = K$. We obtain:

$$\overline{B}_h = \sum_{k=1}^d (\mathsf{I}_k + \mathsf{II}_k),$$

where

$$\begin{aligned}
\mathsf{I}_k &:= \int_{z\in\mathbb{R}^d}\left(\prod_{k'=1}^d K(z_{k'})\right)\rho_k dz, \\
\rho_k &:= (-h_k z_k)^p \int\limits_{0\leq t_p\leq\cdots\leq t_1\leq 1} \left(\partial_k^p f(\overline{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f(\overline{z}_{k-1})\right) dt_{1:p}, \\
\mathsf{II}_k &:= (-h_k)^p \int_{t\in\mathbb{R}} \frac{t^p}{p!}K(t)dt \int_{z_{-k}\in\mathbb{R}^{d-1}} \partial_k^p f(\overline{z}_{k-1})\left(\prod_{k'\neq k} K(z_{k'})\right) dz_{-k}.
\end{aligned}$$

Notice that for $k \notin \mathcal{R}$, $\partial_k^p f(u) = 0$ for any $u \in \mathcal{U}_n(x)$, thus I_k and II_k vanish. Therefore:

$$\overline{B}_h = \sum_{k\in\mathcal{R}} (\mathsf{I}_k + \mathsf{II}_k),$$

Now let us give an equivalent of the bias. First, using Assumption 3, for any $k \in \mathcal{R}$, we can define the modulus of continuity of $\partial_k^p f$ on $\mathcal{U}_n(w)$ by:

$$\Omega_{nk} := \sup_{z,z'\in\,\mathcal{U}_n(w)} \left|\partial_k^p f(z') - \partial_k^p f(z)\right|$$

Then we decompose II_k as follows:

$$\mathsf{II}_k = \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt}{p!} \partial_k^p f(w) + R_k,$$

with $R_k := \frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt}{p!} \int_{z_{-k}\in\mathbb{R}^{d-1}} (\partial_k^p f(\overline{z}_{k-1}) - \partial_k^p f(w)) \left(\prod_{k'\neq k} K(z_{k'}) \right) dz_{-k}$ such that:

$$|R_k| \leq h_k^p \left| \int_{t\in\mathbb{R}} \frac{t^p}{p!} K(t)dt \right| \Omega_{nk} \|K\|_1^{d-1} \tag{33}$$

since $\left|\partial_k^p f(\overline{z}_{k-1}) - \partial_k^p f(w)\right| \leq \Omega_{nk}$.

It remains to bound I_k. From the definition of ρ_k in (50), we write:

$$\begin{aligned} |\rho_k| &\leq |h_k z_k|^p \left| \int_{0\leq t_p \leq \cdots \leq t_1 \leq 1} \left[\partial_j^p f(\overline{z}_{k-1} - t_p h_k z_k e_k) - \partial_j^p f(\overline{z}_{k-1}) \right] dt_{1:p} \right| \\ &\leq |h_k z_k|^p \frac{\Omega_{nk}}{p!}. \end{aligned}$$

Therefore:

$$\begin{aligned} |\mathsf{I}_k| &= \left| \int_{z\in\mathbb{R}^d} \left(\prod_{k'=1}^d K(z_{k'}) \right) \rho_k dz \right| \\ &\leq \frac{h_k^p}{p!} \Omega_{nk} \int_{z\in\mathbb{R}^d} \left| z_k^p \prod_{k'=1}^d K(z_{k'}) \right| dz \\ &\leq \|K\|_1^{d-1} \int_{t\in\mathbb{R}} \left| \frac{t^p}{p!} K(t) \right| dt \times h_k^p \Omega_{nk} \end{aligned} \tag{34}$$

Since $\mathcal{U}_n(w) \underset{n\to\infty}{\longrightarrow} \{w\}$, by continuity of $\partial_k^p f$:

$$\Omega_{nk} \underset{n\to\infty}{\longrightarrow} 0.$$

Therefore for n large enough, combining (33) and (34):

$$|\mathsf{I}_k| + |R_k| \leq \frac{|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!} \max_{k\in\mathcal{R}} |\partial_k^p f(w)| \times h_k^p$$

Therefore, since:

$$\overline{B}_h = \sum_{k\in\mathcal{R}} (\mathsf{II}_k + \mathsf{I}_k) = \sum_{k\in\mathcal{R}} \left(\frac{(-h_k)^p \int_{t\in\mathbb{R}} t^p K(t)dt \; \partial_k^p f(w)}{p!} + R_k + \mathsf{I}_k \right),$$

we obtain:

$$\left|\overline{B}_h\right| \leq \mathsf{C}_{\overline{\text{bias}}} \sum_{k\in\mathcal{R}} h_k^p,$$

with $\mathsf{C}_{\overline{\text{bias}}} := \frac{2|\int_{t\in\mathbb{R}} t^p K(t)dt|}{p!} \max_{k\in\mathcal{R}} |\partial_k^p f(w)|$.

2. We want to apply Bernstein's inequality (cf Lemma 10) to $\bar{f}_h(w)$. We first obtain an almost sure upper bound:

$$\begin{aligned} |\bar{f}_{h1}(w)| &= \frac{1}{f_X(X_1)} \prod_{k=1}^d \frac{\left|K\left(\frac{w_k - W_{1k}}{h_k}\right)\right|}{h_k} \\ &\leq \overline{\mathsf{M}}_h, \end{aligned} \tag{35}$$

where $\overline{\mathsf{M}}_h := \frac{\mathsf{C}_{\bar{\mathsf{M}}}}{\prod_{k=1}^d h_k}$ with $\mathsf{C}_{\bar{\mathsf{M}}} := \frac{\|K\|_\infty^d}{\delta}$.
Then we control the variance:

$$\begin{aligned}
\mathsf{Var}\left(\bar{f}_{h1}(w)\right) &= \mathsf{Var}\left(\frac{1}{f_X(X_1)}\prod_{k=1}^d \frac{K\left(\frac{w_k - W_{1k}}{h_k}\right)}{h_k}\right) \\
&\leq \mathbb{E}\left[\left(\frac{1}{f_X(X_1)}\prod_{k=1}^d \frac{K\left(\frac{w_k - W_{1k}}{h_k}\right)}{h_k}\right)^2\right] \\
&= \int_{u\in\mathbb{R}^d} \left\{\prod_{k=1}^d \frac{1}{h_k^2} K\left(\frac{w_k - u_k}{h_k}\right)^2\right\} \frac{f(u)}{f_X(u_{1:d_1})} du \\
&\leq \frac{1}{\delta \prod_{k=1}^d h_k} \int_{z\in\mathbb{R}^d} \left\{\prod_{k=1}^d K(z_k)^2\right\} f(w - Hz) dz \\
&\leq \bar{\mathsf{v}}_h,
\end{aligned} \tag{36}$$

where $\bar{\mathsf{v}}_h := \frac{\mathsf{C}_\sigma^2}{4\prod_{k=1}^d h_k}$.Therefore we obtain from Bernstein's inequality (cf Lemma 10):

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \leq 2\exp\left(-\min\left(\frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h}, \frac{3n\sigma_h}{4\overline{\mathsf{M}}_h}\right)\right)$$

We compare the rates:

$$\begin{aligned}
& \frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h} \leq \frac{3n\sigma_h}{4\overline{\mathsf{M}}_h} \\
\iff \quad & \mathsf{C}_\sigma \sqrt{\frac{(\log n)^a}{n\prod_{k=1}^d h_k}} = \sigma_h \leq \frac{3\bar{\mathsf{v}}_h}{\overline{\mathsf{M}}_h} = \frac{3\mathsf{C}_\sigma^2}{4\mathsf{C}_{\bar{\mathsf{M}}}} \\
\iff \quad & \prod_{k=1}^d h_k \geq \frac{4\mathsf{C}_{\bar{\mathsf{M}}}^2}{9\mathsf{C}_\sigma^2} \frac{(\log n)^a}{n} \\
\iff \quad & \mathsf{Cond}(h).
\end{aligned}$$

Therefore, if $\mathsf{Cond}(h)$ is satisfied:

$$\mathbb{P}\left(\overline{\mathcal{B}}_h^c\right) \leq 2e^{-\frac{n\sigma_h^2}{4\bar{\mathsf{v}}_h}} = 2e^{-(\log n)^a}.$$

3. We now apply Bernstein's inequality (cf Lemma 10) to $\frac{1}{n}\sum_{i=1}^n |\bar{f}_{hi}(w)|$. From the upper bounds (35) and (36), we obtain:

$$\mathbb{P}\left(\mathcal{B}_{|\bar{f}|h}^c\right) \leq 2\exp\left(-\min\left(\frac{n\mathsf{C}_{\bar{\mathsf{E}}}^2}{4\bar{\mathsf{v}}_h}, \frac{3n\mathsf{C}_{\bar{\mathsf{E}}}}{4\overline{\mathsf{M}}_h}\right)\right).$$

We calculate the rates: by definition of $\bar{\mathsf{v}}_h$ and $\overline{\mathsf{M}}_h$,

$$\frac{n\mathsf{C}_{\bar{\mathsf{E}}}^2}{4\bar{\mathsf{v}}_h} = \frac{\mathsf{C}_{\bar{\mathsf{E}}}^2}{\mathsf{C}_\sigma^2} n \prod_{k=1}^d h_k$$

$$\frac{3n\mathsf{C}_{\bar{\mathsf{E}}}}{4\overline{\mathsf{M}}_h} = \frac{3\mathsf{C}_{\bar{\mathsf{E}}}}{4\mathsf{C}_{\bar{\mathsf{M}}}} n \prod_{k=1}^d h_k$$

Hence:

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \le 2e^{-\mathsf{C}_{\gamma|f|} n \prod_{k=1}^d h_k},$$

with $\mathsf{C}_{\gamma|f|} := \min\left(\frac{\mathsf{C}^2_{\bar{\mathsf{E}}}}{\mathsf{C}^2_\sigma}; \frac{3\mathsf{C}_{\bar{\mathsf{E}}}}{4\mathsf{C}_{\bar{\mathsf{M}}}}\right)$.

5.4.4 Proof of Lemma 6

1. First, we write $\bar{Z}_{hij}$ more explicitly: for any bandwidth h, any observation $i = 1 : n$ and any direction j,

$$\begin{aligned}
\bar{Z}_{hij} &= \frac{\partial}{\partial h_j}\left(\frac{K(\frac{w_j - W_{ij}}{h_j})}{h_j}\right)\frac{\prod\limits_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{f_X(X_i)\prod_{k\neq j} h_k} \\
&= \frac{-\left(K(\frac{w_j - W_{ij}}{h_j}) + \frac{w_j - W_{ij}}{h_j}K'(\frac{w_j - W_{ij}}{h_j})\right)\prod\limits_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{f_X(X_i) h_j \prod_{k=1}^d h_k} \\
&= \frac{-J(\frac{w_j - W_{ij}}{h_j})\prod\limits_{k\neq j} K(\frac{w_k - W_{ik}}{h_k})}{f_X(X_i) h_j \prod_{k=1}^d h_k}
\end{aligned}$$

where we recall $J : \mathbb{R} \to \mathbb{R}$ is the function $t \mapsto tK'(t) + K(t)$.
Note then that the support of J is included in the support of K, and by integration by part, we obtain for any $l \in \mathbb{N}$:

$$\int_{\mathbb{R}} t^l J(t)dt = \int_{\mathbb{R}} t^l (tK(t))' dt = -l \int_{\mathbb{R}} t^l K(t)dt \tag{37}$$

In particular, since K is of order p, for $l = 0 : p-1$, $\int_{\mathbb{R}} t^l J(t)dt = 0$ and $\int_{\mathbb{R}} t^p J(t)dt \neq 0$.
We recall the notation $\cdot$ for the multiplication terms by terms of two vectors. Using Assumption 2, if $j \notin \mathcal{R}$, $f(w - h\cdot z) - f(\tilde{z}_{-j}) = 0$ for any $z \in \mathbb{R}^d$. Thus we obtain:

$$\begin{aligned}
\mathbb{E}\left[\bar{Z}_{h1j}\right] &= -\frac{1}{h_j\prod_{k=1}^d h_k}\int_{u\in\mathbb{R}^d} J(\tfrac{w_j - u_j}{h_j})\left(\prod_{k\neq j} K(\tfrac{w_k - u_k}{h_k})\right) f(u)du \\
&= -\frac{1}{h_j}\int_{z_j\in\mathbb{R}} J(z_j)dz_j \int_{z_{-j}\in\mathbb{R}^{d-1}}\left(\prod_{k\neq j} K(z_k)\right) f(w - h\cdot z)dz_{-j} = 0
\end{aligned}$$

Therefore $\mathbb{E}\left[\bar{Z}_{h1j}\right] = 0$ for $j \notin \mathcal{R}$.
Now, we deal with the case $j \in \mathcal{R}$. Let us fix $j \in \mathcal{R}$. We denote $\tilde{z}_{-j} := w - (Hz)_{-j} = w - \sum\limits_{k\neq j} h_k z_k e_k$ (with $\{e_k\}_{k=1}^d$ the canonic basis of $\mathbb{R}^d$). Then we write:

$$\mathbb{E}\left[\bar{Z}_{h1j}\right] = \frac{-1}{h_j\prod_{k=1}^d h_k}\int_{u_{-j}\in\mathbb{R}^{d-1}}\left(\prod_{k\neq j} K(\tfrac{w_k - u_k}{h_k})\right)\left[\int_{u_j\in\mathbb{R}} J(\tfrac{w_j - u_j}{h_j})du_j - f(\tilde{z}_{-j})\int_{\mathbb{R}} J(z_j)dz_j\right]du_{-j}.$$

Then for fixed $\{z_k\}_{k\neq j}$, denoting $f_j : z_j \mapsto f(w - h\cdot z)$, we apply Lemma 9 with the settings $d' = 1$, $u = \tilde{z}_{-j}$, $h = h_j$, $f_0 = f_j$, $p = p$, $K = J$, then

$$\begin{aligned}
\mathbb{E}\left[\bar{Z}_{h1j}\right] &= \frac{-1}{h_j\prod_{k=1}^d h_k}\int_{u_{-j}\in\mathbb{R}^{d-1}}\left(\prod_{k\neq j} K(\tfrac{w_k - u_k}{h_k})\right)[\mathsf{I}_1 + \mathsf{II}_1]\,du_{-j} \\
&= \widetilde{\mathsf{I}}_j + \widetilde{\mathsf{II}}_j, \qquad (38)
\end{aligned}$$

where

$$\tilde{\mathsf{I}}_j := (-h_j)^{-1} \int_{z\in\mathbb{R}^d} \left(\prod_{k\neq j} K(z_k)\right) J(z_j)\tilde{\rho}_j dz, \tag{39}$$

$$\text{with } \tilde{\rho}_j := (-h_j z_j)^p \int\limits_{0\le t_p\le\cdots\le t_1\le 1} \left(\partial_j^p f(\tilde{z}_{-j} - t_p h_j z_j e_j) - \partial_j^p f(\tilde{z}_{-j})\right) dt_{1:p}, \tag{40}$$

$$\text{and } \widetilde{\mathsf{II}}_j := (-h_j)^{p-1} \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \int_{z_{-j}\in\mathbb{R}^{d-1}} \partial_j^p f(\tilde{z}_{j-1}) \left(\prod_{k'\neq j} K(z_{k'})\right) dz_{-j}.$$

Now let us determine an equivalent of $\mathbb{E}\left[\bar{Z}_{hj}\right]$. For this purpose, let us introduce the modulus of continuity of $\partial_j^p f$ on $\mathcal{U}_n(w)$ (which is well defined by Assumption 3):

$$\Omega_{nj} := \sup_{z,z'\in\,\mathcal{U}_n(w)} \left|\partial_j^p f(z') - \partial_j^p f(z)\right|$$

Then we write:

$$\widetilde{\mathsf{II}}_j = (-h_j)^{p-1} \partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt + \widetilde{R}_j, \tag{41}$$

with

$$\widetilde{R}_j := (-h_j)^{p-1} \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt \int_{z_{-j}\in\mathbb{R}^{d-1}} \left(\partial_j^p f(\tilde{z}_{-j}) - \partial_j^p f(w)\right) \left(\prod_{k\neq j} K(z_k)\right) dz_{-j}.$$

In particular:

$$\begin{aligned} |\widetilde{R}_j| &\le h_j^{p-1} \left|\int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right| \int_{z_{-k}\in\mathbb{R}^{d-1}} \Omega_{nj} \prod_{k\neq j} |K(z_k)| dz_{-j} \\ &= h_j^{p-1} \Omega_{nj} \left|\int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right| \|K\|_1^{d-1} \end{aligned} \tag{42}$$

Now let us bound $\tilde{\mathsf{I}}_j$ defined in (39). First, we bound $\tilde{\rho}_j$, defined in (40):

$$\begin{aligned} |\tilde{\rho}_j| &= (h_j|z_j|)^p \left| \int\limits_{0\le t_p\le\cdots\le t_1\le 1} \left(\partial_j^p f(\tilde{z}_{-j} - t_p h_j z_j e_j) - \partial_j^p f(\tilde{z}_{-j})\right) dt_{1:p} \right| \\ &\le h_j^p |z_j|^p \frac{\Omega_{nj}}{p!}, \end{aligned}$$

which lead to:

$$\begin{aligned} \left|\tilde{\mathsf{I}}_j\right| &= h_j^{-1} \left| \int_{z\in\mathbb{R}^d} \left(\prod_{k\neq j} K(z_k)\right) J(z_j)\tilde{\rho}_j dz \right| \\ &\le h_j^{p-1} \Omega_{nj} \|K\|_1^{d-1} \int_{z_j\in\mathbb{R}} \frac{|z_j|^p}{p!} |J(z_j)| dz_j. \end{aligned} \tag{43}$$

Therefore using (41) then (42) and (43):

$$\begin{aligned} \left|\mathbb{E}\left[\bar{Z}_{h1j}\right]\right| \le \left|\widetilde{\mathsf{II}}_j\right| + \left|\tilde{\mathsf{I}}_j\right| \quad &\le \quad h_j^{p-1} \left|\partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right| + \left|\widetilde{R}_j\right| + \left|\tilde{\mathsf{I}}_j\right| \\ &\le \mathsf{C}_{E\bar{Z},j} h_j^{p-1} + h_j^{p-1} \Omega_{nj} \left(\left|\int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right| \|K\|_1^{d-1} + \|K\|_1^{d-1} \int_{\mathbb{R}} \frac{|t|^p}{p!} |J(t)| dt \right) \end{aligned}$$

with $\mathsf{C}_{E\bar{Z},j} := \left|\partial_j^p f(w) \int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right|$.

Finally, notice that by continuity of $\partial_j^p f$ (Assumption 3), since $\mathcal{U}_n(w) \underset{n\to\infty}{\longrightarrow} \{w\}$:

$$\Omega_{nj} \underset{n\to\infty}{\longrightarrow} 0.$$

Thus for n large enough:

$$\Omega_{nj}\left(\left|\int_{t\in\mathbb{R}} \frac{t^p}{p!} J(t)dt\right| \|K\|_1^{d-1} + \|K\|_1^{d-1} \int_{z_j\in\mathbb{R}} \frac{|z_j|^p}{p!} |J(z_j)| dz_j\right) \le \frac{1}{2}\mathsf{C}_{E\bar{Z},j},$$

which lead to the result (12) of Theorem 2:

$$\frac{1}{2}\mathsf{C}_{E\bar{Z},j} h_j^{p-1} \le \left|\mathbb{E}\left[\bar{Z}_{hj}\right]\right| \le \frac{3}{2}\mathsf{C}_{E\bar{Z},j} h_j^{p-1}.$$

To obtain the result (13) of Theorem 2, just note that:

$$\begin{aligned}
\mathbb{E}\left[|\bar{Z}_{h1j}|\right] &= \frac{1}{h_j \prod_{k=1}^d h_k} \int_{u\in\mathbb{R}^d} \left| J(\tfrac{w_j-u_j}{h_j}) \left(\prod_{k\neq j} K(\tfrac{w_k-u_k}{h_k})\right)\right| f(u)du \\
&= h_j^{-1} \int_{z\in\mathbb{R}^d} \left| J(z_j) \left(\prod_{k\neq j} K(z_k)\right)\right| f(w-Hz)dz \\
&\le \mathsf{C}_{E|\bar{Z}|} h_j^{-1},
\end{aligned}$$

with $\mathsf{C}_{E|\bar{Z}|} := \|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_1 \|K\|_1^{d-1}$.

2. We first bound $\bar{Z}_{hij}$ a.s. and its variance.

$$\begin{aligned}
|\bar{Z}_{hij}| &= \frac{\left|J(\frac{w_j-W_{ij}}{h_j})\right| \prod_{k\neq j} \left|K(\frac{w_k-W_{ik}}{h_k})\right|}{f_X(X_i) h_j \prod_{k=1}^d h_k} \\
&\le \frac{\|J\|_\infty \|K\|_\infty^{d-1}}{\delta h_j \prod_{k=1}^d h_k} = \frac{\mathsf{C}_{\mathsf{M}\bar{Z}}}{h_j \prod_{k=1}^d h_k} =: \mathsf{M}_{\bar{Z},hj} \qquad (44)
\end{aligned}$$

For the variance:

$$\begin{aligned}
\mathsf{Var}\left(\bar{Z}_{hij}\right) &\le \mathbb{E}\left[\bar{Z}_{hij}^2\right] \\
&= \int_{\mathbb{R}^d} J(\tfrac{w_j-u_j}{h_j})^2 \left(\prod_{k\neq j} K(\tfrac{w_k-u_k}{h_k})^2\right) \frac{f_{XY}(u)}{f_X(u_{1:d_1})^2 h_j^2 \prod_{k=1}^d h_k^2} du \\
&= \frac{1}{h_j^2 \prod_{k=1}^d h_k} \int_{\mathbb{R}^d} J(z_j)^2 \left(\prod_{k\neq j} K(z_k)^2\right) \frac{f(w-Hz)}{f_X(x-(Hz)_{1:d_1})} dz \\
&\le \frac{\|f\|_{\infty,\,\mathcal{U}_n(w)} \|J\|_2^2 \|K\|_2^{2(d-1)}}{\delta h_j^2 \prod_{k=1}^d h_k} = \frac{\mathsf{C}_{\mathsf{v}\bar{Z}}}{h_j^2 \prod_{k=1}^d h_k} =: \mathsf{v}_{\bar{Z},hj}. \qquad (45)
\end{aligned}$$

We apply Bernstein's inequality (cf Lemma 10) to $\bar{Z}_{hj}$:

$$\mathbb{P}\left(\mathcal{B}_{\bar{Z},hj}^c\right) \le 2\exp\left(-\min\left(\frac{n(\frac{\lambda_{hj}}{2})^2}{4\mathsf{v}_{\bar{Z},hj}}, \frac{3n\frac{\lambda_{hj}}{2}}{4\mathsf{M}_{\bar{Z},hj}}\right)\right)$$

Let us compare the rates:

$$\begin{aligned}
& \frac{n(\frac{\lambda_{hj}}{2})^2}{4\mathsf{v}_{\bar{Z},hj}} \leq \frac{3n\frac{\lambda_{hj}}{2}}{4\mathsf{M}_{\bar{Z},hj}} \\
\iff \quad & \mathsf{C}_\lambda \sqrt{\frac{(\log n)^a}{nh_j^2 \prod_{k=1}^d h_k}} = \lambda_{hj} \leq \frac{6\mathsf{v}_{\bar{Z},hj}}{\mathsf{M}_{\bar{Z},hj}} = \frac{6\mathsf{C}_{\mathsf{v}\bar{Z}}}{\mathsf{C}_{\mathsf{M}\bar{Z}} h_j} \\
\iff \quad & \prod_{k=1}^d h_k \geq \frac{\mathsf{C}_{\mathsf{M}\bar{Z}}^2 \mathsf{C}_\lambda^2}{6^2 \mathsf{C}_{\mathsf{v}\bar{Z}}^2} \frac{(\log n)^a}{n} \\
\iff \quad & \mathsf{Cond}_{\bar{Z}}(h).
\end{aligned}$$

So, if $\mathsf{Cond}_{\bar{Z}}(h)$ is satisfied:

$$\mathbb{P}\left(\mathcal{B}^c_{\bar{Z},hj}\right) \leq 2e^{\frac{-n(\lambda_{hj}/2)^2}{4\mathsf{v}_{\bar{Z},hj}}} = 2e^{\frac{-\delta}{\|f\|_{\infty,\, \mathcal{U}_n(w)}}(\log n)^a} = 2e^{-\gamma_{Z,n}}$$

3. We apply Bernstein's inequality (cf Lemma 10) to $\frac{1}{n}\sum_{i=1}^n |\bar{Z}_{hij}|$ using the upper bounds (44) and (45):

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{Z}|,h}\right) \leq 2\exp\left(-\min\left(\frac{n(\mathsf{C}_{E|\bar{Z}|}h_j^{-1})^2}{4\mathsf{v}_{\bar{Z},hj}}, \frac{3n\mathsf{C}_{E|\bar{Z}|}h_j^{-1}}{4\mathsf{M}_{\bar{Z},hj}}\right)\right)$$

Let us calculate the rate: by definition of $\mathsf{C}_{\mathsf{v}\bar{Z}}hj$ and $\mathsf{M}_{\bar{Z},hj}$,

$$\begin{aligned}
\frac{n(\mathsf{C}_{E|\bar{Z}|}h_j^{-1})^2}{4\mathsf{v}_{\bar{Z},hj}} &= \frac{\mathsf{C}_{E|\bar{Z}|}^2}{4\mathsf{C}_{\mathsf{v}\bar{Z}}} n \prod_{k=1}^d h_k \\
\frac{3n\mathsf{C}_{E|\bar{Z}|}h_j^{-1}}{4\mathsf{M}_{\bar{Z},hj}} &= \frac{3\mathsf{C}_{E|\bar{Z}|}}{4\mathsf{C}_{\mathsf{M}\bar{Z}}} n \prod_{k=1}^d h_k
\end{aligned}$$

Hence:

$$\mathbb{P}\left(\mathcal{B}^c_{|\bar{f}|h}\right) \leq 2e^{-\mathsf{C}_{\gamma|\bar{Z}|} n \prod_{k=1}^d h_k},$$

with $\mathsf{C}_{\gamma|\bar{Z}|} := \min\left(\frac{\mathsf{C}_{E|\bar{Z}|}^2}{4\mathsf{C}_{\mathsf{v}\bar{Z}}}; \frac{3\mathsf{C}_{E|\bar{Z}|}}{4\mathsf{C}_{\mathsf{M}\bar{Z}}}\right)$.

5.4.5 Proof of Lemma 7

1. We decompose $\Delta_{Z,hj}$ as follows:

$$\Delta_{Z,hj} := Z_{hj} - \bar{Z}_{hj} = \frac{1}{n}\sum_{i=1}^n \left(\frac{f_X(X_i) - \tilde{f}_X(X_i)}{\tilde{f}_X(X_i)}\right)\bar{Z}_{hij}.$$

Using $\bar{Z}_{hij} = 0$ when $X_i \notin \mathcal{U}_h(x)$:

$$|\Delta_{Z,hj}| \leq \left\|\frac{f_X - \tilde{f}_X}{\tilde{f}_X}\right\|_{\infty,\, \mathcal{U}_n(x)} \frac{1}{n}\sum_{i=1}^n \left|\bar{Z}_{hij}\right|. \tag{46}$$

First we deal with $\left\|\frac{f_X - \tilde{f}_X}{\tilde{f}_X}\right\|_{\infty,\, \mathcal{U}_n(x)}$. By definition of $\widetilde{A}_n$:

$$\mathbb{1}_{\widetilde{A}_n}\left\|\frac{f_X - \tilde{f}_X}{\tilde{f}_X}\right\|_{\infty,\, \mathcal{U}_n(x)} \leq M_X\left(\frac{(\log n)^d}{n}\right)^{1/2}, \tag{47}$$

Now let us give an upper bound of $\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right|$. Using Lemma 6,

$$\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}}\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right| \leq \mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\bar{Z}_{hij}\right| - \mathbb{E}\left[\left|\bar{Z}_{h1j}\right|\right]\right| + \mathbb{E}\left[\left|\bar{Z}_{h1j}\right|\right]$$
$$\leq 2\mathsf{C}_{E|\bar{Z}|}h_j^{-1}$$

To conclude, combining this last result with (47) and (46):

$$\mathbb{1}_{\mathcal{B}_{|\bar{Z}|,hj}\cap\widetilde{A}_n}\left|\Delta_{Z,hj}\right| \leq 2\mathsf{C}_{E|\bar{Z}|}M_X h_j^{-1}\left(\frac{(\log n)^d}{n}\right)^{1/2}$$
$$\leq \frac{2\mathsf{C}_{E|\bar{Z}|}M_X}{\mathsf{C}_\lambda(\log n)^{\frac{a}{2}}}\lambda_{hj} = \frac{\mathsf{C}_{\mathsf{M}\Delta Z}}{(\log n)^{\frac{a}{2}}}\lambda_{hj},$$

since $\prod_{k=1}^{d} h_k \leq h_0^d = \frac{1}{(\log n)^d}$.

2. We decompose Δ_h as follows:

$$\Delta_h := \hat{f}_h(w) - \bar{f}_h(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{f_X(X_i) - \tilde{f}_X(X_i)}{\tilde{f}_X(X_i)}\right)\bar{f}_{hi}(w).$$

Using $\bar{f}_{hi}(w) = 0$ when $X_i \notin \mathcal{U}_h(x)$:

$$|\Delta_h| \leq \left\|\frac{f_X - \tilde{f}_X}{\tilde{f}_X}\right\|_{\infty,\,\mathcal{U}_n(x)}\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right|.$$

We have proved in (47): $\mathbb{1}_{\widetilde{A}_n}\left\|\frac{f_X - \tilde{f}_X}{\tilde{f}_X}\right\|_{\infty,\,\mathcal{U}_n(x)} \leq M_X\left(\frac{(\log n)^d}{n}\right)^{1/2}$.

Let us now give an upper bound of $\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right|$. Using Lemma 5,

$$\mathbb{1}_{\mathcal{B}_{|\bar{f}|h}}\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right| \leq \mathbb{1}_{\mathcal{B}_{|\bar{f}|h}}\left|\frac{1}{n}\sum_{i=1}^{n}\left|\bar{f}_{hi}(w)\right| - \mathbb{E}\left[\left|\bar{f}_{hi}(w)\right|\right]\right| + \mathbb{E}\left[\left|\bar{f}_{h1}(w)\right|\right]$$
$$\leq 2\mathsf{C}_{\bar{\mathsf{E}}}.$$

Therefore:

$$\mathbb{1}_{\widetilde{A}_n\cap\mathcal{B}_{|\bar{f}|h}}\left|\Delta_h\right| \leq 2\mathsf{C}_{\bar{\mathsf{E}}}M_X\left(\frac{(\log n)^d}{n}\right)^{1/2} \leq \frac{2\mathsf{C}_{\bar{\mathsf{E}}}M_X}{\mathsf{C}_\sigma}(\log n)^{-\frac{a}{2}}\sigma_h,$$

since $\prod_{k=1}^{d} h_k \leq h_0^d = (\log n)^{-d}$.

5.4.6 Proof of Lemma 9

We first denote

$$B := \int_{\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'} h_j^{-1}K\left(\frac{u_j - u'_j}{h_j}\right)\right)f_0(u')du' - f_0(u)\int_{\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'}K\left(z_j\right)\right)dz.$$

Then we obtain by integration by parts:

$$B := \int_{z\in\mathbb{R}^{d'}}\left(\prod_{j=1}^{d'}K(z_j)\right)(f_0(u - h\cdot z) - f_0(u))dz \tag{48}$$

For any $z \in \mathbb{R}^{d'}$, we denote $\overline{z}_0 := w$ and for $k = 1 : d'$, $\overline{z}_k := u - \sum_{j=1}^{k} h_j z_j e_j$ (where $\{e_j\}_{j=1}^{d'}$ is the canonical basis of $\mathbb{R}^{d'}$). Then, we write:

$$f_0(u - h \cdot z) - f_0(u) = \sum_{k=1}^{d'} f_0(\overline{z}_k) - f_0(\overline{z}_{k-1}) \tag{49}$$

Then we apply Taylor's theorem (cf Lemma 11) to the functions $g_k : t \in [0,1] \mapsto f_0(\overline{z}_{k-1} - t h_k z_k e_k)$, $k \in (1 : d')$:

$$f_0(\overline{z}_k) - f_0(\overline{z}_{k-1}) = g_k(1) - g_k(0) = \sum_{l=1}^{p} \frac{(-z_k h_k)^l}{l!} \partial_k^l f_0(\overline{z}_{k-1}) + \rho_k,$$

where we denote for short:

$$\rho_k := \rho_k(z,h,u) = (-h_k z_k)^p \int_{0 \le t_p \le \cdots \le t_1 \le 1} \left(\partial_k^p f_0(\overline{z}_{k-1} - t_p h_k z_k e_k) - \partial_k^p f_0(\overline{z}_{k-1}) \right) dt_{1:p}. \tag{50}$$

We introduce the notation

$$\mathsf{I}_k := \int_{z \in \mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j) \right) \rho_k dz$$

and for any $z \in \mathbb{R}^{d'}$, we denote $z_{-k} \in \mathbb{R}^{d'-1}$ the vector z without its k^{th} variable, then (48) becomes:

$$\begin{aligned} B &= \int_{z \in \mathbb{R}^{d'}} \left(\prod_{j=1}^{d'} K(z_j) \right) \left(\sum_{k=1}^{d'} \sum_{l=1}^{p} \frac{(-h_k)^l}{l!} z_k^l \partial_k^l f_0(\overline{z}_{k-1}) + \rho_k \right) dz \\ &= \sum_{k=1}^{d'} \left(\mathsf{I}_k + \sum_{l=1}^{p} \frac{(-h_k)^l}{l!} \int_{z_{-k} \in \mathbb{R}^{d-1}} \partial_k^l f_0(\overline{z}_{k-1}) \left(\prod_{j \ne k} K(z_j) \right) \int_{z_k \in \mathbb{R}} z_k^l K(z_k) dz_k dz_{-k} \right) \end{aligned}$$

Since K has at least $p-1$ zero moments, the terms with $l \le p-1$ vanish, leading to:

$$\begin{aligned} B &= \sum_{k=1}^{d'} \left(\mathsf{I}_k + \frac{(-h_k)^p \int_{t \in \mathbb{R}} t^p K(t) dt}{p!} \int_{z_{-k} \in \mathbb{R}^{d'-1}} \partial_k^p f_0(\overline{z}_{k-1}) \left(\prod_{j \ne k} K_j(z_j) \right) dz_{-k} \right) \\ &=: \sum_{k=1}^{d'} (\mathsf{I}_k + \mathsf{II}_k), \end{aligned} \tag{51}$$

with $\mathsf{II}_k := (-h_k)^p \int_{t \in \mathbb{R}} \frac{t^p}{p!} K(t) dt \int_{z_{-k} \in \mathbb{R}^{d'-1}} \partial_k^p f_0(\overline{z}_{k-1}) \left(\prod_{j \ne k} K(z_j) \right) dz_{-k}$.

Acknowledgement.

The author is extremely grateful to Claire Lacour and Vincent Rivoirard for suggesting me to study this problem, for the stimulating discussions, the helpful advices and the careful proofreading.

References

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Comput. Statist. Data Anal.*, 36(3):279–298.

Beaumont, M., Zhang, W., and Balding, D. (2002). Approximate bayesian computation in population genetics. *Genetics*, 162(4):2025–2035.

Bertin, K., Lacour, C., and Rivoirard, V. (2016). Adaptive pointwise estimation of conditional density function. *Ann. Inst. H. Poincaré Probab. Statist.*, 52(2):939–980.

Biau, G., Cérou, F., and Guyader, A. (2015). New insights into approximate bayesian computation. *Ann. Inst. H. Poincaré Probab. Statist.*, 51(1):376–403.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Donoho, D. L. and Low, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.*, 20(2):944–970.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer Science & Business Media.

Efromovich, S. (2007). Conditional density estimation in a regression setting. *Ann. Statist.*, 35(6):2504–2535.

Efromovich, S. (2010a). Dimension reduction and adaptation in conditional density estimation. *Journal of the American Statistical Association*, 105(490):761–774.

Efromovich, S. (2010b). Oracle inequality for conditional density estimation and an actuarial example. *Ann. Inst. Statist. Math.*, 62(2):249–275.

Fan, J., Yao, Q., and Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika*, 91(4):819–834.

Fan, J.-q., Peng, L., Yao, Q.-w., and Zhang, W.-y. (2009). Approximating conditional density functions using dimension reduction. *Acta Mathematicae Applicatae Sinica, English Series*, 25(3):445–456.

Faugeras, O. P. (2009). A quantile-copula approach to conditional density estimation. *J. Multivariate Anal.*, 100(9):2083–2099.

Fernández-Soto, A., Lanzetta, K., Chen, H.-W., Levine, B., and Yahata, N. (2002). Error analysis of the photometric redshift technique. *Monthly Notices of the Royal Astronomical Society*, 330(4):889–894.

Goldenshluger, A. and Lepski, O. (2011). Bandwidth selection in kernel density estimation: oracle inequalities and adaptive minimax optimality. *Ann. Statist.*, 39(3):1608–1632.

Györfi, L. and Kohler, M. (2007). Nonparametric estimation of conditional distributions. *IEEE Trans. Inform. Theory*, 53(5):1872–1879.

Hall, P., Racine, J., and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, 99(468):1015–1026.

Holmes, M. P., Gray, A. G., and Isbell, C. L. (2010). Fast kernel conditional density estimation: A dual-tree monte carlo approach. *Computational Statistics & Data Analysis*, 54(7):1707 – 1718.

Hyndman, R. J., Bashtannyk, D. M., and Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *J. Comput. Graph. Statist.*, 5(4):315–336.

Hyndman, R. J. and Yao, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *J. Nonparametr. Stat.*, 14(3):259–278.

Izbicki, R. and Lee, A. B. (2016). Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316.

Izbicki, R. and Lee, A. B. (2017). Converting high-dimensional regression to high-dimensional conditional density estimation. *Electron. J. Statist.*, 11(2):2800–2831.

Jeon, J. and Taylor, J. W. (2012). Using conditional kernel density estimation for wind power density forecasting. *J. Amer. Statist. Assoc.*, 107(497):66–79.

Lafferty, J. and Wasserman, L. (2008). Rodeo: Sparse, greedy nonparametric regression. *Ann. Statist.*, 36(1):28–63.

Liu, H., Lafferty, J. D., and Wasserman, L. A. (2007). Sparse nonparametric density estimation in high dimensions using the rodeo. In *International Conference on Artificial Intelligence and Statistics*, pages 283–290.

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. (2012). Approximate bayesian computation methods. *Statistics and Computing*, 22(6):1167–1180.

Otneim, H. and Tjøstheim, D. (2017). Conditional density estimation using the local gaussian correlation. *Statistics and Computing*, pages 1–19.

Rosenblatt, M. (1969). Conditional probability density and regression estimators. In *Multivariate Analysis, II (Proc. Second Internat. Sympos., Dayton, Ohio, 1968)*, pages 25–31. Academic Press, New York.

Sart, M. (2017). Estimating the conditional density by histogram type estimators and model selection. *ESAIM: Probability and Statistics*, 21:34–55.

Takeuchi, I., Le, Q. V., Sears, T. D., and Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7(Jul):1231–1264.

Takeuchi, I., Nomura, K., and Kanamori, T. (2009). Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Comput.*, 21(2):533–559.

Wasserman, L. and Lafferty, J. D. (2006). Rodeo: Sparse nonparametric regression in high dimensions. In *Advances in Neural Information Processing Systems*, pages 707–714.